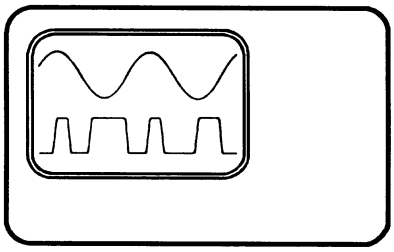
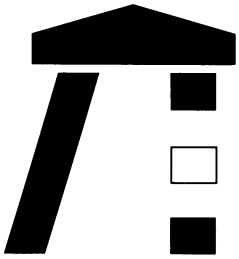


FACHBEREICH
ELEKTROTECHNIK UND
INFORMATIONSTECHNIK

Psychoakustische Signalverbesserung und Geräuschreduktion in Kraftfahrzeugen

vorgelegt von
Dipl.-Ing. Mike Peters



UNIVERSITÄT
KAISERSLAUTERN

Psychoakustische Signalverbesserung und Geräuschreduktion in Kraftfahrzeugen

Vom Fachbereich Elektrotechnik und Informationstechnik
der Universität Kaiserslautern
zur Verleihung des akademischen Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)
genehmigte Dissertation

von

Dipl.-Ing. Mike Peters

aus Feldafing

D 386

Tag der mündlichen Prüfung: 12. April 2002

Dekan des Fachbereiches: Prof. Dr.-Ing. R. Urbansky

Promotionskomission:
Vorsitzender: Prof. Dr.-Ing. A. Potchinkov
1. Berichterstatter: Prof. Dr.-Ing. (em.) W. Rupprecht
2. Berichterstatter: Prof. Dr.-Ing. R. Urbansky

The machine does not isolate us
from the great problems
of nature
but plunges us
more deeply into them.

Antoine de Saint-Exupéry

Vorwort

Die vorliegende Arbeit entstand während meiner Tätigkeit im Forschungs- und Ingenieurzentrum der BMW AG. Viele Ideen und Anregungen ergaben sich bei der Entwicklungsarbeit im Bereich Sprachsignalverarbeitung, Multimedia- und Infotainmentsysteme oder stammen aus den Erfahrungen, die im Entwicklungsprozeß gewonnen wurden.

Besonderer Dank gilt Herrn Prof. Dr.-Ing. W. Rupprecht für die fortwährende Unterstützung dieser Arbeit und die wertvollen Ratschläge. Ebenso danke ich Herrn Prof. Dr.-Ing. R. Urbansky für die Übernahme des Korreferats. Herrn Dr.-Ing. W. Sauer-Greff danke ich für die vielen fachlichen Hinweise. Die Herren Dipl.-Ing. W. Weishaupt und Dr.-Ing. U. Weinmann haben diese Arbeit in der BMW AG ermöglicht. Vielen Dank.

Für die themenübergreifenden Diskussionen danke ich den Herren Dipl.-Inf. S. Gäßner, Dipl.-Ing. A. Dittrich und Dipl.-Ing. J. Eckert. Außerdem bedanke ich mich bei meinen Kollegen in der BMW AG und im Institut für Nachrichtentechnik der Universität Kaiserslautern und bei allen meinen Freunden, die auf verschiedende Weise zum Fortgang und Gelingen dieser Arbeit beitrugen.

Zum Schluß danke ich Dir, Julia. Du hast mich motiviert und unterstützt, diese Arbeit neben meiner beruflichen Tätigkeit bei BMW an den wenigen freien Tagen fertigzustellen.

München, im August 2001 *Mike Peters*

Inhaltsverzeichnis

Abkürzungen und Formelzeichen

Skalare Größen und Funktionen werden in Standarddruck, Variablen werden *kursiv* dargestellt. Kleinbuchstaben sind vornehmlich für Größen im Zeitbereich und Großbuchstaben für den Frequenzbereich bestimmt. Vektoren und Matrizen werden fett gedruckt. Speziell verwendete Begriffe oder Verfahren werden *kursiv* dargestellt. Kontinuierliche Signale im Zeitbereich werden in Abhängigkeit von der Zeit t und im Frequenzbereich als Funktion der Frequenz f dargestellt. Zeitdiskrete Signale werden durch k im Zeitbereich und n im Frequenzbereich indiziert. Teilweise erfolgt im Frequenzbereich auch die Darstellung anhand der normierten Frequenz Ω. Kurzzeitdarstellungen werden durch den Fensterindex m ausgedrückt.

Formelzeichen

t, f	kontinuierliche Zeit und Frequenz
k, n, m	diskreter Zeit- und Frequenzindex, Fensterindex
$n(k)$, $s(k)$, $x(k)$	abgetastetes Nutz- und Störsignal sowie Signalgemisch aus beiden Anteilen im Zeitbereich
$N(n)$, $S(n)$, $X(n)$	abgetastete Beträge des Nutz-, Stör- und Mischsignalspektrums
Ω	normierte Frequenz
z_k	Bark Index, Tonheit
$r_{xx}(k)$	Autokorrelationsfunktion des Signals x
$r_{xy}(k)$	Kreuzkorrelationsfunktion der Signale x und y
$R_{xx}(k)$	(Auto)Leistungsdichte des Signals x
$R_{xy}(k)$	Kreuzleistungsdichte der Signale x und y
$\tilde{R}_{xx}(n, m)$	Kurzzeitleistungsdichte im m-ten Zeitrahmen
$E\{\dots\}$	Erwartungswert

$H(k),\ G(k)$	Übertragungsfunktion, Gewichtsfunktion
$\bar{s}(k)$	Mittelwert eines Signals oder Wertes
$\hat{s}(k)$	Schätzung eines Signals
$g(k),\ h(k)$	Impulsantwort
$H(z)$	Z-Transformierte
$w(k)$	Fensterfunktion
$S(n, m)$	Fouriertransformierte des Signals s im m-ten Fenster mit dem Frequenzindex n
p_e	Schalldruck mit Bezugsgröße
I_s	psychoakustische Intensität des Signals s
I_M	Intensität des Maskierers
L_T	Maskierschwelle
L_M	Pegel des Maskierers
C_{xy}	Kohärenz zwischen den Signalen x und y
f_a	Abtastfrequenz
$\nabla f(t)$	Gradient der Funktion $f(t)$
$\xi(\Omega)$	a-priori Signal-zu-Störverhältnis
$\psi(\Omega)$	a-posteriori Signal-zu-Störverhältnis
T	Periodendauer (Umlaufzeit)
T_a	Abtastintervall
Δ	Differenz oder Abweichung
a	Oversubtraction-Faktor
b	Spectral Floor
γ	Nichtlinearer Glättungsoperator
$\frac{\partial^n}{\partial x^n} f(x)$	n-te partielle Ableitung von $f(x)$ nach x
q	Quantisierung
c	Schallgeschwindigkeit
π	3.14159
$\Xi(n)$	Preemphasize Filter

M	Anzahl der Fenster
ρ_0	Dichte der Luft
A	(Querschnitts-)Fläche
Z	Schallfeldimpedanz
P_e	Fehlerleistung, relativer Fehler

Abkürzungen

SNR	Signal-zu-Rauschverhältnis
LDS	Leistungsdichtespektrum
LMS	Least-Mean-Square
FIR	Finite Impulse Response (Filter)
IIR	Infinite Impulse Response (Filter)
MFCC	Mel-Frequency-Cepstral-Coefficients
LPC	Linear Predictive Coefficients
AKF, KKF	Auto- und Kreuzkorrelationsfunktion
SNRE	Signalverbesserung
NR	Noise Reduction
MOS	Mean Opinion Score
AI	Artikulationsindex
DFT	Diskrete Fouriertransformation
IDFT	Inverse Diskrete Fouriertransformation
MMSE	Minimum Mean Square Estimation

1 Einleitung

Bei der Übertragung von Sprachsignalen werden die Sprachqualität und die Sprachverständlichkeit vielfach durch Störungen beeinträchtigt. In digitalen Übertragungssystemen ist grundsätzlich mit folgenden Störungsarten zu rechnen:

- akustische Hintergrundstörungen, wie z.B. Straßenlärm, zusätzliche Sprecher
- Rückkopplungen und Echos, z.B. durch Lautsprecher in Freisprecheinrichtungen, Leitungsechos
- Störungen infolge Digitalisierung, z.B. durch Quantisierung und Codierung
- Übertragungsfehler auf dem Kanal, z.B. in Mobilfunksystemen

Offensichtlich unterscheiden sich diese Störungsarten so voneinander, daß zur Verbesserung gestörter Sprachsignale verschiedene Verfahren angewendet werden müssen. In dieser Arbeit werden ausschließlich Maßnahmen zur Verminderung akustischer Hintergrundstörungen mittels Geräuschreduktionssystemen behandelt. Der Einsatz solcher Systeme hat mit der rasanten Verbreitung von Mobiltelefonen und der Nutzung von Freisprecheinrichtungen im Kraftfahrzeug große Bedeutung erlangt. Die beiden Begriffe Störung und Geräusch werden als synonyme Bezeichnungen, unabhängig vom Charakter der Störquelle, verwendet. Die Störsignale sind meist nicht von deterministischer Natur, weswegen sie als stochastische Zufallsprozesse angesehen werden. Im allgemeinen Fall kann nicht ausgeschlossen werden, daß die Störungen beliebig instationär und die akustischen Übertragungswege beliebig zeitvariant sind. Um dennoch das Problem der Geräuschreduktion zu handhaben, sollen nachfolgende Voraussetzungen gelten:

- die akustischen Übertragungswege seien linear
- die Übertragungsfunktionen seien zeitinvariant oder höchstens langsam zeitveränderlich
- alle Geräuschquellen sollen sich in einer einzigen äquivalenten Geräuschquelle konzentrieren lassen, abweichende Annahmen werden explizit erläutert
- Sprach- und Störsignal sollen sich additiv überlagern
- Sprach- und Störsignal seinen physikalisch und damit auch statistisch unabhängig
- Sprach- und Störsignal sollen sich zumindestens kurzzeitig als stationäre und ergodische Zufallsprozesse beschreiben lassen

Der Einsatz von Freisprecheinrichtungen bei der Sprachkommunikation in Fahrzeugen erfordert die Reduktion der mit dem Sprachsignal erfaßten Umgebungsgeräusche. Die akustischen Störungen beeinträchtigen in der Regel die Verständlichkeit des zu übertragenden Sprachsignals.

In der Literatur wurden zahlreiche Verfahren und Ansätze zur Geräuschreduktion vorgeschlagen und beschrieben. Prinzipiell können diese Ansätze in drei Kategorien unterteilt werden: Einkanalige Geräuschreduktionssysteme, wie zum Beispiel das Verfahren der Spektralen Subtraktion, mehrkanalige Geräuschkompensationsverfahren, die mindestens ein Störgeräusch-Referenzsignal benötigen, und adaptive Mikrophonarrays, die zur Erfassung des Sprachsignals ein richtungsselektives Reduktionsverfahren (*beam forming*) einsetzen.

Diese Arbeit fokussiert ausschließlich auf das Problem der einkanaligen Geräuschreduktionssysteme, wie sie häufig in Kraftfahrzeugen oder Telefonen aus Kosten- und konstruktiven Gründen zu finden sind. Mehrkanalige Verfahren werden nur der Vollständigkeit halber am Rande behandelt.

Einkanalige Verfahren sind durch den Kompromiß zwischen der Dämpfung der störenden Geräusche und den unvermeidbaren Verzerrungen des Sprachsignals und der verbleibenden Restörungen gekennzeichnet. Diese Verzerrungen sind als sporadisch auftretende, tonartige Restörungen (*musical tones*) bzw. als Verfärbungen des Sprachsignals wahrnehmbar. Solche Fehler im Ausgangssignal werden wegen ihrer tonalen Struktur als äußerst störend empfunden und verschlechtern den subjektiven Höreindruck.

In letzter Zeit sind deshalb Verfahren mit dem Ziel entwickelt worden, möglichst alle auftretenden Verzerrungen zu unterdrücken. So wurden zum Beispiel nichtlineare Methoden, bekannt aus der Bildverarbeitung, oder spezielle Detektionsalgorithmen entworfen, um das Problem geschlossen zu lösen.

Besonders neu sind Verfahren, die psychoakustische Eigenschaften des menschlichen Gehörs nutzen, um wenigstens einen Teil der auftretenden Verzerrungen zu *verdecken*. So kommen hier Methoden zum Einsatz, die durch Formulierung einer psychoakustischen Gewichtungsregel einen Kompromiß zwischen Höhe der Geräuschdämpfung und der resultierenden Sprachverständlichkeit eingehen.

In der vorliegenden Arbeit diente ein klassisches einkanaliges Geräuschreduktionsverfahren als Ausgangsbasis für die Entwicklung eines neuen psychoakustisch-parametrischen Verfahrens. Dabei wurde von Modellen der Spracherzeugung und Wahrnehmung der menschlichen Sprache ausgegangen, um geeignete Methoden für die psychoakustische Geräuschreduktion und Signalverbesserung zu finden. Das Ergebnis sind drei neue Verfahren, die sich je nach Eingangssignal adaptiv auf die Charakteristik des Gehörs einstellen und dabei Verzerrungen

des Sprachsignals und der Reststörung unterhalb der psychoakustischen Wahrnehmbarkeitsschwelle, der sogenannten *Mithörschwelle*, halten. Das führt zu einer spürbaren Verbesserung des subjektiven Höreindrucks und hat positiven Einfluß auf die Sprachverständlichkeit. In wesentlichen Bestandteilen dieser Arbeit werden Aspekte der psychologischen Wahrnehmung akustischer Signale und bekannte psychoakustische Eigenschaften des menschlichen Gehörs für die auditive Signalverbesserung, Geräuschreduktion und die Identifikation akustischer Systeme ausgenutzt. Dementsprechend wird im ersten Teil eine kurze Einführung in die Theorie der Signalverarbeitung und Psychoakustik gegeben. Daran anschließend folgt die Vorstellung eines Verfahrens zur auditiven Signalverbesserung und Geräuschreduktion unter Ausnutzung psychoakustischer Verdeckungseffekte. Dieser Abschnitt ist besonders ausführlich gestaltet, da er den Hauptbestandteil der Arbeit bildet. Der dritte Teil erläutert experimentelle Untersuchungen und die Bewertung der verschiedenen Verfahren. Abschließend folgen Zusammenfassung und ein wissenschaftlicher Ausblick.

2 Theoretische Grundlagen

Kenntnisse der spektralen wie der statistischen Beschreibung des Sprach- und Störsignals sind für die Sprachverarbeitung unabdingbar. Das gilt nicht nur wegen des Allgemeinverständnisses, sondern auch für den unmittelbaren Einsatz in den beschriebenen Verarbeitungsmethoden. Im folgenden sind einige Grundbegriffe und Verfahren zusammengestellt, soweit sie später benötigt werden. Besonderes Augenmerk wird dabei auf die Bestimmung spektraler oder statistischer Charakteristika der Signale gerichtet.

2.1 Zeitkontinuierliche Signale

Zur Beschreibung des physikalischen Sprachsignals wird oft die elektrische Repräsentation des Signals $x(t)$ verwendet. Dabei wird angenommen, daß ein ideales Mikrophon den Sprachschall ohne lineare Verfärbungen oder nichtlineare Verzerrungen exakt in ein elektrisches Signal $x(t)$ wandelt, so daß ein idealer Lautsprecher mit dem elektrischen Signal $x(t)$ den ursprünglichen Sprachschall genau reproduzieren kann. Für die Analyse des Sprachsignals wird deshalb zunächst von seiner elektrischen Repräsentation, dem zeitkontinuierlichen Signal $x(t)$ ausgegangen.

Im Raum der zeitkontinuierlichen Energiefunktionen $L_2(\Re)$ sei ein Skalarprodukt

$$\langle x(t), \varphi(t) \rangle = \int_{-\infty}^{\infty} x(t) \cdot \varphi(t) dt, \quad \text{mit} \quad x(t) \in L_2(\Re) \tag{2.1}$$

definiert. Mit diesem Skalarprodukt werden Signale $x(t)$ in den *Bildbereich* transformiert. Die Funktion $\varphi(t)$ wird als Transformationskern bezeichnet. Ist $\varphi(t) = e^{-j\omega t}$, so liegt die bekannte Fouriertransformation vor:

$$X(\omega) = F\{x(t)\} = \int_{-\infty}^{\infty} x(t) \cdot e^{-j\omega t} dt. \tag{2.2}$$

Die Methode der Fourier-Transformation ist in der oben beschriebenen Form nicht universell brauchbar, weil das uneigentliche Integral in Gleichung (2.2) im mathematischen Sinne für zahlreiche wichtige Signaltypen nicht existiert. Beispielsweise konvergiert das Integral schon für stationäre Sinusschwingungen nicht. Berücksichtigt man nur kausale Signale, die erst bei $t = 0$ einsetzen und für $t < 0$ gleich null sind, ist das Konvergenzproblem für $t \to -\infty$ gelöst. Hinsichtlich der Konvergenz für $t \to +\infty$ bietet die einseitige *Laplace*-Transformation[1] mittels einer Dämpfungsfunktion der Form $e^{-\sigma t}$, mit $\sigma > 0$, einen Ansatz zur Lösung des Konvergenzproblems. Die einseitige Laplace-Transformation ist folgendermaßen definiert:

$$X(s) = \int_0^\infty x(t) e^{-st} dt = L\{x(t)\}. \tag{2.3}$$

Dabei wird die komplexe Variable $s = \sigma + j\omega$ eingeführt. Die Dämpfungsfunktion beeinträchtigt nicht die mathematisch korrekte Repräsentation des Signals $x(t)$ im Frequenzbereich, bewirkt aber, daß der Grenzwert von $s(t)$ für $t \to +\infty$ verschwindet, so daß die Transformierte im mathematischen Sinne tatsächlich existiert.

2.2 Zeitdiskrete Signale

Jedes bandbegrenzte Signal mit der Maximalfrequenz f_{max} kann mit der Frequenz $f_a = 1/\mathrm{T}_a$ unter Beachtung des Abtasttheorems $f_a \geq 2 f_{max}$ abgetastet und, abhängig von der Definition von f_{max}, mit Alias-Fehler rekonstruiert werden. Durch die äquidistante Abtastung des zeitkontinuierlichen Signals $x(t)$ entsteht das zeitdiskrete Signal

$$x(k) = x(k \cdot T_a). \tag{2.4}$$

Ist ein Signal nach (2.4) absolut summierbar, d.h. es gilt

$$\sum_{k=-\infty}^{+\infty} |x(k)| < \infty, \tag{2.5}$$

ist die Fouriertransformierte (FT) durch

[1] Bei der zweiseitigen Laplace-Transformation wird auf die Voraussetzung eine kausalen Signals verzichtet und die Konvergenz für $t \to -\infty$ ebenfalls mit Hilfe einer Dämpfungsfunktion erzwungen.

$$x(k) \xrightarrow{FT} X(e^{j\Omega}) = \sum_{k=-\infty}^{+\infty} x(k) e^{-jk\Omega} = F\{x(k)\} \tag{2.6}$$

definiert. Die Umkehrtransformation

$$X(e^{j\Omega}) \xrightarrow{FT^{-1}} x(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\Omega}) e^{jk\Omega} d\Omega = F^{-1}\{X(e^{j\Omega})\} \tag{2.7}$$

stellt die Fouriersynthese dar. Das Spektrum eines abgetasteten Signals ist periodisch mit der Periode f_a. Durch die Einführung der normierten Frequenz Ω mit

$$\Omega = 2\pi \frac{f}{f_a} = \frac{2\pi}{N} \tag{2.8}$$

wird die Grundperiode des Spektrums $X(e^{j\Omega})$ auf das Intervall $\langle 0, 2\pi \rangle$ projiziert. Dabei ist N eine natürliche Zahl.

Tastet man das kontinuierliche Fourierspektrum $X(e^{j\Omega})$ einer endlich langen Folge $x(k)$ mit $k = 0, 1, \ldots, K-1$ exakt in den Frequenzpunkten

$$\Omega_n = \Omega n \tag{2.9}$$

ab, ergibt sich die *Diskrete Fouriertransformierte* (*DFT*) $X(n)$ mit

$$X(n) = \begin{cases} \sum_{k=0}^{K-1} x(k) e^{-j\frac{2\pi}{N}kn} & \text{für } n = 0, 1, \ldots, K\text{-}1 \\ 0 & \text{sonst.} \end{cases} \tag{2.10}$$

Die *Inverse Diskrete Fouriertransformierte* (*IDFT*) ist gegeben durch:

$$x(k) = \begin{cases} \frac{1}{K} \cdot \sum_{n=0}^{K-1} X(n) e^{j\frac{2\pi}{K}kn} & \text{für } n = 0, 1, \ldots, K\text{-}1 \\ 0 & \text{sonst.} \end{cases} \tag{2.11}$$

Eine Erweiterung zu Gleichung (2.6) stellt die sogenannte *z-Transformation* dar. Sie erlaubt die Spektraldarstellung von vielen Signalen, deren Fouriertransformation rein mathematisch nicht existieren. Die zweiseitige z-Transformation (ZT) und die inverse z-Transformation sind folgendermaßen definiert:

$$x(k) \xrightarrow{ZT} X(z) = \sum_{k=-\infty}^{+\infty} x(k) z^{-k} = Z\{x(k)\} \tag{2.12}$$

$$X(z) \xrightarrow{ZT^{-1}} x(k) = \frac{1}{2\pi j} \oint X(z) z^{k-1} dz = ZT^{-1}\{X(z)\}. \tag{2.13}$$

Wenn sowohl $X(e^{j\Omega})$ nach Gleichung (2.6) als auch $X(z)$ nach (2.12) existieren, so gilt auf dem Einheitskreis der komplexen z-Ebene

$$ZT\{x(k)\}_{z=e^{j\Omega}} = F\{x(k)\} = X(e^{j\Omega}) \tag{2.14}$$

und (2.13) geht bei Integration auf dem Einheitskreis in (2.7) über.

2.3 Stochastische Prozesse

Ausgehend von einem beliebigen Zufallsexperiment lassen sich grundlegende Begriffe der statistischen Signaltheorie definieren. Ein meßbares Ergebnis eines Zufallsexperiments wird als *Ereignis* A und die Menge aller möglichen Ereignisse eines Zufallsexperiments wird als *Ereignisfeld* $\boldsymbol{A}$ bezeichnet. Mit $P(A)$ wird die Wahrscheinlichkeit des Ereignisses A aus dem Ereignisfeld $\boldsymbol{A}$ ausgedrückt. Die statistische Signaltheorie benutzt den Zufallsprozeß als Modell für eine Schar von Signalen. Betrachtet man den Zufallsprozess für einen festen Zeitpunkt, so erhält man eine Zufallsvariable. Durch eindeutige Abbildung der Ereignismenge $\boldsymbol{A}$ auf die Menge der reellen Zahlen ergibt sich die sogenannte reelle *Zufallsvariable* $\boldsymbol{x}(A)$. Ähnlich wie nun eine Zufallsvariable $\boldsymbol{x}(A)$ jedem Ereignis aus $\boldsymbol{A}$ eine Zahl x zuordnet, weist ein stochastischer Prozeß $\boldsymbol{x}(A, k)$ jedem Ereignis A aus dem Ereignisfeld $\boldsymbol{A}$ eine Funktion $x(A, k)$ mit der diskreten Zeitvariablen k zu. Die diskrete Zeitfunktion $x(A, k)$ wird als Realisierung des stochastischen Prozesses oder als *Musterfunktion* bezeichnet. Zur Vereinfachung erfolgt die Darstellung der Musterfunktion $x(A, k)$ nachfolgend durch $x(k)$ als stochastisches Signal. Die Menge aller stochastischen Signale $x(k)$ (Musterfunktionen) beschreiben den stochastischen Prozeß $\boldsymbol{x}(A, k)$.

Ein stochastischer Prozeß heißt *stationär*, wenn seine statistischen Eigenschaften invariant gegenüber Verschiebungen der Zeit k sind. Sprach- und Störsignale sind üblicherwiese als stochastische Prozesse anzusehen. Die Prozeßeigenschaften ändern sich mit der Zeit. Die dadurch ausgedrückte Instationarität ist für diese Signale charakteristisch. Gerade in den Variationen der Signalcharakteristika steckt die Information. Ändern sich die statistischen Eigenschaften der Signale im Vergleich zu ihren eigentlichen Zeitverläufen relativ langsam, kann man von Kurzzeit-Stationarität sprechen. Für Sprachsignale gelten diese Aussagen, wenn die Analyse innerhalb von Abschnitten mit 10...20 ms Dauer vorgenommen wird. Von *Stationarität* wird im folgenden meist ausgegangen. Auf Ergänzungen, insbesondere durch die Einführung eines zeitabhängigen Fensterindex m, wird gesondert eingegangen.

2.3.1 Analyse instationärer Signale

Die Fourier-Transformierte hängt vom Parameter $\omega = 2\pi f$ ab, die Abhängigkeit von der Zeit geht verloren. Man erkennt nicht, wann die Spektralanteile des Signals auftreten. Dies ist für die Analyse nichtstationärer Signale unbefriedigend.

Abhilfe bietet die sogenannte Kurzzeit-Fourier-Transformierte (*Short-Time-Fourier-Transform - STFT*), die sich wie folgt berechnen läßt:

$$F\{x(t), \omega, \tau\} = \left\langle x(t), w(t-\tau) \cdot e^{j\omega t} \right\rangle \tag{2.15}$$

und neben dem Frequenzparameter ω noch den Zeitparameter τ besitzt. In den Tranformationskern von (2.15) geht die Fensterfunktion $w(t-\tau)$ ein, die dafür sorgt, daß immer nur ein zeitlicher Ausschnitt von $s(t)$ in dem mit τ gleitendem Analysefenster betrachtet wird. In der Literatur sind verschiedene Fensterfunktionen bekannt, die unterschiedliche Eigenschaften bezüglich des Bandbreite-Zeitprodukts haben, so z.B. das *Hamming-*, *Hanning-* und *Gaborfenster*. Das Spektrum wird ebenfalls nur mit endlicher Auflösung erfaßt (Unschärferelation). Die Aufteilung des Zeit-Bandbreite-Produktes bei verschiedenen Frequenzen ist im Falle der STFT bei allen Werten von τ und ω gleich (konstante Auflösung). Wählt man dagegen als Transformationskern $\varphi(t)$ in (2.1)

$$\varphi(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-\tau}{a}\right), \tag{2.16}$$

erreicht man ein multiples Auflösungsvermögen und es ergibt sich die Wavelet-Transformation (WT) mit dem Mutter-Wavelet $\psi(t)$. Der Parameter τ in (2.16) stellt eine Zeitverschie-

bung, der Parameter a eine Skalierung der Zeitvariablen t dar. Durch entsprechende Wahl des Parameters a kann die Zeitauflösung auf Kosten der Frequenzauflösung erhöht werden. Die Wavelet-Transformation bildet zunächst $x(t)$ in die *Zeit-Skalierungs-Ebene* ab. Wegen der Bandpaßcharakteristik des Wavelets $\psi(t)$ kommt dies aber einer Abbildung in die *Zeit-Frequenz-Ebene* gleich [55].

Das Sprachsignal stellt einen instationären stochastischen Prozeß dar. Nur für kurze Zeitabschnitte (Fenster) endlicher Länge kann das Signal als stationär angesehen werden. Die Analyse des Sprachsignals wird daher in kurzen Signalintervallen, üblicherweise mit einer Dauer von 10...20 ms, vorgenommen. Durch Einsatz einer diskreten Fensterfunktion $w(k)$ der Länge M ergibt sich damit die abschnittsweise diskrete Fourier-Transformierte (WDFT) des Signals $s(k)$. Oftmals wird noch ein *Overlap* eingeführt, so daß sich die Analyserahmen zeitlich überlappen. Die Fensterfunktion und der Overlap verringern den Einfluß der Abschnittsanalyse. Im folgenden werden Signale in halbüberlappende Abschnitte mit jeweils $K = 256$ Abtastwerten[1] getrennt und analysiert.

Um Einflüsse durch die abschnittsweise Analyse des Signals zu vermeiden, verwendet man z.B. die *Hanning*-Fensterfunktion $w(k)$ mit:

$$w(k) = 0.5 - 0.5 \cdot \cos\left(\frac{2\pi k}{K-1}\right), \text{ wobei } k = 0,1,\ldots,K-1. \tag{2.17}$$

Die Kurzzeitanalyse von $s(k)$ ergibt dann für den m-ten Zeitrahmen durch Fensterung mit $w(k)$ und den Overlap p die diskrete Kurzzeit-Fouriertransformierte $S(n, m)$, wobei:

$$S(n,m) = \sum_{k=mp}^{mp+K-1} s(k)w(k)e^{-\frac{j2\pi kn}{K}}. \tag{2.18}$$

Wird ein Overlap $p = 0.5$ gewählt, überlappen sich die Hanningfenster zu 50% und ergeben summiert über die Zeit stets eins. Dadurch wird sichergestellt, daß eine spätere Rekonstruktion des Signals durch das einfache *Overlap-and-Add*-Verfahren problemlos möglich ist.

Die Form der Darstellung in Gleichung (2.18) wird als *Kurzzeit-DFT* und $S(n, m)$ das Kurzzeitspektrum im m-ten Zeitrahmen bezeichnet. Die Kurzzeitdarstellung der z-Transformation ergibt sich dementsprechend als

[1] Bei einer Abtastfrequenz von f_a=11025 Hz entspricht dies etwa einer Signaldauer von 23 ms. Während dieser Zeit kann das Sprachsignal als quasistationär betrachtet werden

$$S(z,m) = \sum_{k=mp}^{mp+K-1} s(k)w(k)z^{-k}. \tag{2.19}$$

2.3.2 Verteilung und Verteilungsdichte

Zentrale Rollen bei der Analyse und Beschreibung stochastischer Prozesse nehmen die *Verteilungsfunktion* und die *Verteilungsdichtefunktion* ein. Die Verteilungsfunktion $P_x(u,k)$ gibt nach (2.20) die Wahrscheinlichkeit P für $x(k) \leq u$ an, wobei u und später auch v als wählbare Schranken zu verstehen sind. Es gilt:

$$P_x(u,k) = P\{x(k) \leq u\}. \tag{2.20}$$

Die Verteilungsdichtefunktion wird bei der Charakterisierung eines stochastischen Signals $x(k)$ mit $p_x(u,k)$ bezeichnet und ist als Ableitung der Verteilungsfunktion $P_x(u,k)$ wie folgt definiert:

$$p_x(u,k) = \frac{\partial P_x(u,k)}{\partial u}. \tag{2.21}$$

Werden zwei Signale $x_{1,2}(k)$ im Zusammenhang behandelt, so ergeben sich die sogenannte Verbundverteilungsdichtefunktion und die Verbundverteilungsfunktion zu:

$$p_{x_1,x_2}(u,v,k_1,k_2) = \frac{\partial^2 P_{x_1,x_2}(u,v,k_1,k_2)}{\partial u \partial v} \tag{2.22}$$

mit

$$P_{x_1,x_2}(u,v,k_1,k_2) = P\{(x_1(k_1) \leq u) \cap (x_2(k_2) \leq v)\}. \tag{2.23}$$

Geht man nun von Stationarität aus, so verschwindet die Abhängigkeit von k in Gleichung (2.21) und in Gleichung (2.22) bleibt anstelle der Abhängigkeit von zwei Beobachtungszeitpunkten nur die Abhängigkeit von deren zeitlichen Abstand

$$\lambda = k_2 - k_1. \tag{2.24}$$

2.3.3 Ergodizität, Erwartungswert, Varianz und Korrelation

Im Falle eines instationären Prozesses $\boldsymbol{x}(A, k)$ hängen die Prozeßkenngrößen *Erwartungswert* und *Varianz* vom Beobachtungszeitpunkt k ab. Allgemein sind nur Scharmittelwerte repräsentativ für einen Zufallsprozeß. Ein Zeitmittelwert sagt dagegen nur etwas über die Musterfunktion aus, für die er berechnet wurde. Er kann für jede Musterfunktion verschieden sein. Es gibt jedoch eine Klasse von stationären Zufallsprozessen, bei denen Scharmittelwerte und Zeitmittelwerte gleich sind. Diese Prozesse nennt man ergodische Prozesse. Stationarität ist damit in jedem Fall Voraussetzung für *Ergodizität*.

Der Erwartungswert eines ergodischen Prozesses $\boldsymbol{x}(A, k)$ mit der eindimensionalen Musterfunktion $x(k)$ ist folgendermaßen definiert:

$$E\{\boldsymbol{x}(A,k)\} = E\{x(k)\} = \int_{-\infty}^{\infty} x(k)\, p_x[x(k)]dx. \tag{2.25}$$

Für die Dichtefunktion einer diskreten Zufallsvariablen $x(A)$ kann man schreiben:

$$p_x[x(A)] = x(A)\cdot\left(\sum_{i=1}^{N} P_x\left[x(A) = x_i\right]\cdot\delta[x(A) - x_i]\right), \tag{2.26}$$

wobei N die Anzahl der Elemente des Ereignisfeldes $\boldsymbol{A}$ und δ den Einheitsimpuls bezeichnen. Somit folgt für den *Erwartungswert* μ_x eines ergodischen Prozesses $\boldsymbol{x}(A, k)$ mit der Musterfunktion $x(k)$ und mit $P_x[x(A) = x_i] = p_i$:

$$E\{\boldsymbol{x}(A,k)\} = E\{x(k)\} = \int_{x=-\infty}^{\infty} x(k)\cdot\left(\sum_{i=1}^{N} p_i\cdot\delta[x(A) - x_i]\right)\cdot dx = \mu_x. \tag{2.27}$$

Die *Varianz* σ_x^2 eines diskreten ergodischen Zufallsprozesses $\boldsymbol{x}(A, k)$ berechnet sich zu:

$$E\left\{\left[\boldsymbol{x}(A,k) - \mu_x\right]^2\right\} = \sigma_x^2. \tag{2.28}$$

Die zweidimensionale Erweiterung von Gleichung (2.25) liefert für ergodische Prozesse die sogenannte *Kreuzkorrelationsfolge* (KKF) zu $x_1(k)$ sowie $x_2(k)$ und drückt deren Ähnlichkeit bei relativer Verschiebung um λ aus. Es folgt für die Kreuzkorrelationsfolge r_{x1x2}:

$$E\{x_1(k)x_2(k+\lambda)\} = r_{x_1x_2}(\lambda) \tag{2.29}$$

und entsprechend gilt für die *Autokorrelationsfolge* (AKF):

$$E\{x(k)x(k+\lambda)\} = r_{xx}(\lambda). \tag{2.30}$$

Aus der Schar der Realisierungen eines Zufallsprozesses ist in aller Regel nur eine einzelne Realisierung für eine Messung verfügbar. Prinzipiell können Korrelationsfunktionen nur für stationäre Prozesse gemessen werden. Außerdem muß die Messung in endlicher Zeit erfolgen, d.h. es steht nur die Schätzung der Kurzzeitkorrelationsfunktion zur Verfügung. Bei zeitdiskreten Zufallsprozessen $x(k)$, die im m-ten Zeitfenster der Länge N analysiert werden, lautet die Meßvorschrift für die Kurzzeit-Autokorrelationsfolge dann:

$$\hat{r}_{xx}(\lambda, m) = \frac{1}{N}\sum_{k=0}^{N-1} x(k,m)x(k+\lambda,m). \tag{2.31}$$

Die Kurzzeit-Kreuzkorrelationsfunktion $\hat{r}_{xy}(\lambda, m)$ zweier stochastischer Signale $x(k, m)$ und $y(k, m)$ der Länge N im m-ten Zeitrahmen läßt sich folgendermaßen bestimmen:

$$\hat{r}_{xy}(\lambda, m) = \frac{1}{N}\sum_{k=0}^{N-1} x(k,m)y(k+\lambda,m). \tag{2.32}$$

Während die AKF eine symmetrische Folge ergibt, ist die KKF im allgemeinen nicht symmetrisch. Beide Folgen haben die Länge $2N-1$.

2.3.4 Unkorreliertheit, Orthogonalität und statistische Unabhängigkeit

Zwei über derselben Ereignismenge definierte Signale $x(k)$ und $y(k)$ sind *unkorreliert*, wenn für sie gilt:

$$E\{x(k)y(k)\} = E\{x(k)\}\cdot E\{y(k)\}. \tag{2.33}$$

Sie werden als *orthogonal* bezeichnet, wenn die KKF $E\{x(k)y(k+\lambda)\}$ für alle λ verschwindet, d.h. es gilt:

$$E\{x(k)y(k+\lambda)\} = 0. \tag{2.34}$$

Ein Vergleich beider Definitionen läßt erkennen, daß zwei unkorrelierte Zufallsvariablen auch orthogonal sind, wenn mindestens eine davon mittelwertfrei ist.

Zwei Zufallsvariablen $x(A) = x$ und $y(A) = y$ sind dann statistisch *unabhängig*, wenn für ihre gemeinsame Wahrscheinlichkeitsdichtefunktion gilt:

$$p_{xy}(x, y) = p_x(x)\,p_y(y). \tag{2.35}$$

Dies bedeutet, daß zwischen zwei statistisch unabhängigen Größen, beispielsweise dem Abtastwert des Sprachsignals und dem Abtastwert einer Störung, kein Zusammenhang besteht. Im Umkehrschluß ermöglicht die Kenntnis des Wertes der einen Variablen keinen Rückschluß auf den Wert der anderen Variablen.

2.3.5 Spektrale Leistungsdichte

Die Fourier- oder z-Transformation eines zeitlich unbegrenzten Zufallsignals läßt sich nicht geschlossen berechnen, da im allgemeinen keine Konvergenz erzielbar ist. Zu einer sinnvollen Spektralbeschreibung gelangt man erst nach einem Übergang zu Korrelationsfolgen und Leistungsdichtespektren. Das *Autoleistungsdichtespektrum* $R_{xx}(\omega)$ eines stationären Signals $x(t)$ ist die Fouriertransformierte der Autokorrelationsfunktion $r_{xx}(\tau)$. Es gilt für zeitkontinuierliche Signale:

$$R_{xx}(\omega) = \int_{-\infty}^{\infty} r_{xx}(\tau) e^{-j\omega\tau} d\tau. \tag{2.36}$$

Diese Berechnungsvorschrift ist als *Wiener-Khintchine-Theorem* bekannt. Voraussetzung für seine Gültigkeit ist, daß die AKF des stochastischen Signals schneller mit der Zeit τ abklingt als die Funktion $1/(|\tau|)$, vgl. [129].

Für zeitdiskrete Signale $x(k)$ mit $k = 0, 1, \ldots, N-1$ ergibt sich das Leistungsdichtespektrum $R_{xx}(n)$ als diskrete Fouriertransformierte der Autokorrelationsfolge $r_{xx}(\lambda)$ zu:

$$R_{xx}(n) = \sum_{\lambda=-(N-1)}^{N-1} r_{xx}(\lambda) \cdot e^{-j2\pi\frac{\lambda}{N}n}. \tag{2.37}$$

Liegt im m-ten Fenster der Länge N ein quasistationäres Signal $x(k, m)$ vor, kann eine *Schätzung* für das LDS $\hat{R}_{xx}(n, m)$ durch DFT der geschätzten Autokorrelationsfolge $\hat{r}_{xx}(\lambda, m)$ aus Gleichung (2.31) gewonnen werden.

Mit der geschätzten AKF $\hat{r}_{xx}(\lambda, m)$ ergibt sich für das geschätzte diskrete Leistungsdichtespektrum $\hat{R}_{xx}(n, m)$ des Prozesses $x(k, m)$ im m-ten Signalrahmen:

$$\hat{R}_{xx}(n,m) = \sum_{\lambda=-(N-1)}^{N-1} \hat{r}_{xx}(\lambda,m) \cdot e^{-j2\pi\frac{\lambda}{N}n}. \tag{2.38}$$

Analog läßt sich das *Kreuzleistungsdichtespektrum* $R_{xy}(\omega)$ für zeitkontinuierliche Zufallsprozesse und $R_{xy}(n)$ für zeitdiskrete Prozesse durch Fouriertransformation der Kreuzkorrelation $r_{xy}(\tau)$ bzw. $r_{xy}(\lambda)$ bestimmen. Es gilt dann für die Schätzung der Kreuzleistungsdichte im m-ten Zeitrahmen:

$$\hat{R}_{xy}(n,m) = \sum_{\lambda=-(N-1)}^{N-1} \hat{r}_{xy}(\lambda,m) \cdot e^{-j\pi\frac{\lambda}{N}n}. \tag{2.39}$$

Die Kreuzkorrelationsfolge $\hat{r}_{xy}(\lambda, m)$ der Länge $2N-1$ nach Gleichung (2.32) ist im allgemeinen nicht symmetrisch. Es liegt nahe, ein zeitbegrenztes Signalstück $x(k, m)$ mit $k = 0, 1, \ldots, N-1$ unmittelbar einer DFT nach (2.18) zu unterziehen und das Kurzzeit-Leistungsdichtespektrum im m-ten Rahmen durch Betragsquadrieren der Fouriertransformierten zu schätzen. Das sogenannte *Periodogramm* ergibt sich zu

$$\begin{aligned} \left|X(n,m)\right|^2 &= DFT\left\{x(k,m)\right\} \cdot DFT^*\left\{x(k,m)\right\} \\ &= DFT\left\{x(k,m)\right\} \cdot DFT\left\{x(N-k,m)\right\}. \end{aligned} \tag{2.40}$$

Die Inverse des Periodogramms gemäß (2.40) im m-ten Signalfenster erhält man wegen der impliziten Periodizität der IDFT, siehe Abschnitt 2.4, zu:

$$\begin{aligned} \tilde{r}_{xx}(u,m) &= DFT^{-1}\left\{\left|X(n,m)\right|^2\right\} = \frac{1}{N}\sum_{u=0}^{N-1}\left|X(n,m)\right|^2 e^{j\frac{2\pi u}{N}n} \\ &= \sum_{k=0}^{N-1} x(k) \cdot x(k+u)_{\bmod N}. \end{aligned} \tag{2.41}$$

Durch Vergleich mit Gleichung (2.31) wird deutlich, daß die „zyklische“ Schätzung $\tilde{r}_{xx}(\lambda)$ der Autokorrelationsfunktion $r_{xx}(\lambda)$ bis auf den Faktor $1/N$ mit der „linearen“ Schätzung $\hat{r}_{xx}(\lambda)$ übereinstimmt, wenn das Signalfenster $w(k)$ so gewählt wird, daß $x(k, m)$ auf $N/2$ begrenzt und dann mit Nullen für $k = (N/2) + 1, (N/2) + 2, \ldots, N - 1$ aufgefüllt wird.

2.3.6 Kohärenzfunktion

Zur Beschreibung der frequenzabhängigen Korrelation zweier stationärer Signale $x(k)$ und $y(k)$ eignet sich die komplexwertige *Kohärenzfunktion* $C_{xy}(\Omega)$ bzw. deren Betragsquadrat. Die komplexe Kohärenzfunktion ist nach [31] folgendermaßen definiert:

$$C_{xy}(\Omega) = \frac{\left|R_{xy}(\Omega)\right|^2}{R_{xx}(\Omega) \cdot R_{yy}(\Omega)}. \tag{2.42}$$

Die Kohärenzfunktion kann nur Werte mit $0 \le C_{xy}(\Omega) \le 1$ annehmen [45]. Sind die Signale $x(k)$ und $y(k)$ unkorreliert, so ergibt sich $C_{xy}(\Omega) = 0$. Ist $y(k)$ das Ausgangssignal eines linearen, zeitinvarianten Systems $h(k)$ mit der Anregung $x(k)$, so gilt mit (2.54) und (2.55)

$$C_{xy}(\Omega) = \frac{\left|R_{xx}(\Omega) \cdot H(\Omega)\right|^2}{R_{xx}(\Omega) \cdot \left[R_{xx}(\Omega) \cdot \left|H(\Omega)\right|^2 \right]} = 1. \tag{2.43}$$

2.4 Lineare zeitinvariante Systeme

Die in folgenden Abschnitten behandelten Signale und Systeme sind als stochastisch anzusehen, das heißt, sie sind weitgehend zeitvariant. Zunächst wird aber angenommen, daß sich die statistischen Eigenschaften der Signale und die Impulsantworten der linearen Systeme im Zeitverlauf nicht ändern. So wie das zeitkontinuierliche, lineare, zeitinvariante Übertragungssystem durch die Antwort $h(t)$ auf einen Dirac-Impuls $\delta(t)$ vollständig charakterisiert ist, so ist das zeitdiskrete, lineare, zeitinvariante Übertragungssystem durch die Anwort $h(k)$ auf eine Eins-Impulsfolge $\delta(k)$ vollständig beschrieben. Die vorgestellten Spektraltransformationen lassen sich unter den in Abschnitt 2.2 getroffenen Voraussetzungen auf allgemeine zeitdiskrete Signale $s(k)$ wie auch auf die *Impulsantwort* $h(k)$ linearer, zeitinvarianter, diskreter Systeme anwenden. Die z-Transformierte der Impulsanwort wird als *Übertragungsfunktion* $H(z)$ mit

$$H(z) = \mathrm{ZT}\{h(k)\} \tag{2.44}$$

bezeichnet. Die diskrete Fouriertransformierte $H(e^{j\Omega})$ von $h(k)$ mit

$$H(e^{j\Omega}) = \mathrm{FT}\{h(k)\} \tag{2.45}$$

gibt den Frequenzgang des Systems an, wobei $|H(e^{j\Omega})|$ als Betragsfrequenzgang und $arg[H(e^{j\Omega})]$ als Phasengang bezeichnet werden. Mit diesen systembeschreibenden Ausdrükken lassen sich lineare, zeitinvariante Systeme mit Hilfe der Faltung und der bekannten Faltungssätze angeben:

$$y(k) = x(k) * h(k) = \sum_{u=-\infty}^{+\infty} x(u)h(k-u), \tag{2.46}$$

$$Y(z) = X(z)H(z) \tag{2.47}$$

und

$$Y(e^{j\Omega}) = X(e^{j\Omega})H(e^{j\Omega}). \tag{2.48}$$

Die inversen Transformationen von (2.47) und (2.48) zeigen die Analogie dieser drei Formeln. Der entsprechende Faltungssatz für die DFT

$$Y(n) = H(n)X(n) = DFT\{h(k)\}DFT\{x(k)\} \tag{2.49}$$

besagt allerdings wegen der impliziten DFT-Periodizität, daß

$$\begin{aligned} y(k) &= DFT^{-1}\{Y(n)\} = x(k) \otimes h(k) = h(k) \otimes x(k) \\ &= \sum_{i=0}^{N} x(i)h([k-i]_{\mathrm{mod}\,N}) \end{aligned} \tag{2.50}$$

gilt. Das endliche Signal $x(k)$ wird mit der N-periodisch wiederholten Impulsantwort *zyklisch* gefaltet. Die entstehende Folge $y(k)$ stimmt im allgemeinen nicht mit dem Ergebnis der linearen Faltung überein. Die Autokorrelationsfolge $r_{xx}(\lambda)$ des Eingangssignals $x(k)$ wird durch

das lineare, zeitinvariante System mit der Impulsantwort $h(k)$ folgendermaßen abgebildet. Es ergibt sich somit:

$$r_{xx}(\lambda) * h(\lambda) = \sum_{u=-\infty}^{\infty} h(u) \cdot E\{x(k) \cdot x(k+\lambda-u)\}. \tag{2.51}$$

Durch Vertauschung von Faltungssumme und der Erwartungswertbildung und mit einfacher Variablensubstitution findet man:

$$\begin{aligned} r_{xx}(\lambda) * h(\lambda) &= \sum_{u=-\infty}^{\infty} h(u) \cdot E\{x(k)x(x+\lambda-u)\} \\ &= r_{xy}(\lambda). \end{aligned} \tag{2.52}$$

Durch Faltung mit der inversen Impulsantwort des Übertragungssystems $h(-\lambda)$ ergibt sich weiter mit (2.52)

$$r_{xx}(\lambda) * h(\lambda) * h(-\lambda) = r_{yy}(\lambda). \tag{2.53}$$

Ein wichtiger Unterschied zwischen den beiden Beziehungen in (2.52) und (2.53) wird nach Fouriertransformation deutlich. Mit (2.37) ergibt sich für das Kreuzleistungsdichtespektrum:

$$R_{xy}(n) = R_{xx}(n) \cdot H(n) \tag{2.54}$$

und mit Gleichung (2.39) folgt für das Autoleistungsdichtespektrum des Signals $y(k)$:

$$R_{yy}(n) = R_{xx}(n) \cdot |H(n)|^2 . \tag{2.55}$$

2.5 Orthogonalitätsprinzip

Die lineare Differenzengleichung

$$y(k) = a_1 y(k-1) + \ldots + a_n y(k-m) = b_1 s(k-1) + \ldots + b_m s(k-n) \tag{2.56}$$

beschreibt ein zeitinvariantes, diskretes System, dessen Eingangssignal $s(k)$ und Ausgangssignal $y(k)$, wie im oberen Teil der Abbildung 2.1 dargestellt, zusammenhängen. Dabei soll zunächst die Zeitvarianz der Systemparameter unbeachtet bleiben.

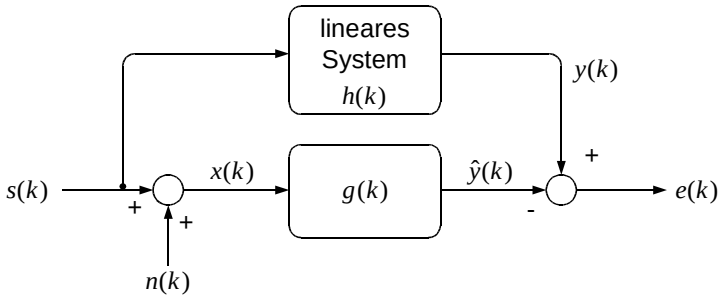


Abbildung 2.1 Optimierung eines linearen Systems

Allgemein läßt sich der Zusammenhang aus (2.56) in die Form

$$y(k) = \sum_{v=0}^{\infty} h(v)s(k-v) \tag{2.57}$$

bringen. Dabei bezeichnet $h(v)$ in (2.57) die Impulsanwort des Systems, dessen Übertragungsfunktion $H(z)$ durch die einseitige *z-Transformierte* der Impulsantwort $h(v)$ als:

$$H(z) = \sum_{v=0}^{\infty} h(v)z^{-v} \tag{2.58}$$

gegeben ist. Das System mit der Impulsantwort $g(k)$ im unteren Teil der Abbildung 2.1 wird als Schätzsystem verstanden, das eine Schätzung $\hat{y}$ für den tatsächlichen Wert y liefert. Der Schätzfehler berechnet sich nach Abbildung 2.1 als:

$$e(k) = y(k) - \hat{y}(k). \tag{2.59}$$

Minimiert man nun den mittleren quadratischen Fehler $E\{e^2(k)\}$ durch Veränderung der Koeffizienten des Schätzsystems $g(k)$, so erhält man die Koeffizienten des optimalen Schätzsystems oder kurz des Optimalfilters $g_{\text{opt}}(k)$. Für den mittleren quadratischen Fehler ergibt sich dabei:

$$E\{e^2(k)\} = E\left\{\left[y(k) - \hat{y}(k)\right]^2\right\} \rightarrow \min. \tag{2.60}$$

Mit dem Ansatz aus (2.57) gilt für ein nichtkausales, optimales Filter

$$\hat{y}_{\text{opt}}(k) = \sum_{u=-\infty}^{\infty} g_{\text{opt}}(u)x(k-u). \tag{2.61}$$

Bezeichnet man nun den gefundenen minimalen Fehler mit

$$e_{\min}(k) = y(k) - \hat{y}_{\text{opt}}(k) \tag{2.62}$$

und den mittleren quadratischen Fehler des nichtoptimalen Systems mit

$$\begin{aligned} E\{e^2(k)\} &= E\left\{\left[y(k) - \hat{y}_{\text{opt}}(k) - \alpha\hat{y}_{\Delta}(k)\right]^2\right\} \\ &= E\left\{\left[y(k) - \sum_{u=-\infty}^{\infty} x(k-u)g_{\text{opt}}(u) - \alpha\sum_{u=-\infty}^{\infty} x(k-u)g_{\Delta}(u)\right]^2\right\} \end{aligned} \tag{2.63}$$

und bedenkt man, daß stets

$$E\{e^2(k)\} - E\{e_{\min}^2(k)\} \geq 0 \tag{2.64}$$

gelten muß, dann läßt sich (2.64) mit (2.63) und (2.62) wie folgt darstellen:

$$E\{e^2(k)\} - E\{e_{\min}^2(k)\} = -2\alpha E\{e_{\min}(k)\hat{y}_{\Delta}(k)\} + \alpha^2 E\{\hat{y}_{\Delta}^2\} \geq 0. \tag{2.65}$$

Dabei stellen $g_{\Delta}(u)$ die Abweichung vom optimalen Schätzsystem $g_{\text{opt}}(u)$ und α ein beliebiger skalarer Faktor dar.

Diese Ungleichung kann aber nur dann für beliebige α gelten, wenn der lineare Term in (2.65) verschwindet. Durch Nullsetzen des linearen Terms erhält man:

$$\begin{aligned} 0 &= E\{e_{\min}(k)\hat{y}_{\Delta}(k)\} \\ &= E\{e_{\min} \cdot \sum_{u=-\infty}^{\infty} x(k-u)g_{\Delta}(u)\}. \end{aligned} \tag{2.66}$$

Vertauscht man nun in (2.66) Erwartungswertbildung und Summation, ergibt sich das sogenannte *Orthogonalitätsprinzip* und die Bedingung für Optimalität des Schätzfilters:

$$E\{e_{\min}(k)x(k-u)\} = 0, \quad \text{für alle } u \text{ mit} \quad g_{\Delta}(u) \neq 0. \tag{2.67}$$

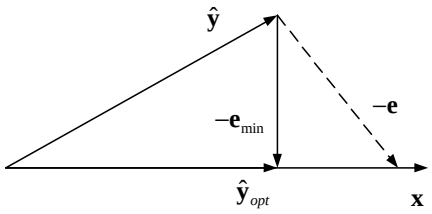


Abbildung 2.2 Geometrische Interpretation des Orthonalitätsprinzips

Der minimale Fehler $e_{\min}(k)$ ist zu beliebigen Abtastwerten des Eingangssignals $x(k)$ orthogonal. Die Orthogonalität zwischen dem minimalen Fehler $e_{\min}(k)$ und dem Ausgang des optimalen Filters $\ddot{y}_{\text{opt}}(k)$ kann genauso gezeigt werden. Das Orthogonalitätsprinzip läßt eine geometrische Interpretation nach Abbildung 2.2 zu: Ein Vektor

$$\mathbf{y} = \left[y(0), y(1), \ldots, y(K-1) \right] \tag{2.68}$$

ist durch einen Vektor

$$\hat{\mathbf{y}}_{\text{opt}} = \left[\hat{y}_{\text{opt}}(0), \hat{y}_{\text{opt}}(1), \ldots, \hat{y}_{\text{opt}}(K-1) \right], \tag{2.69}$$

der parallel zu einem Vektor

$$\mathbf{x} = [x(0), x(1), \ldots, x(K-1)] \tag{2.70}$$

verläuft, anzunähern. Die Länge des Fehlervektors $|\mathbf{e}|$ mit

$$\mathbf{e} = [e(0), e(1), \ldots, e(K-1)] \tag{2.71}$$

wird minimal, wenn dieser *orthogonal* zu $\mathbf{x}$ und $\hat{\mathbf{y}}_{\text{opt}}$ steht, siehe Abbildung 2.2.

2.6 Optimalfilter

2.6.1 Nichtkausales Wiener Filter

Wendet man das Orthogonalitätsprinzip durch Einsetzen von (2.59) und (2.61) in (2.67) an, so ergibt sich:

$$E\{[y(k) - \sum_{v=-\infty}^{\infty} x(k-v) g_{\text{opt}}(v)] x(k-u)\} = 0. \tag{2.72}$$

Mit der allgemeinen Kreuzkorrelation

$$r_{pq}(u) = E\{p(k) q(k+u)\} \tag{2.73}$$

folgt dann nach Vertauschen der Reihenfolge von Summation und Erwartungswertbildung in der Beziehung (2.72):

$$r_{xy}(u) - \sum_{v=-\infty}^{\infty} g_{\text{opt}}(v) r_{xx}(u-v) = 0. \tag{2.74}$$

Diese Gleichung bezeichnet man als diskrete *Wiener-Hopf-Gleichung*. Sie stellt die implizite Bestimmungsgleichung für die optimalen Filterkoeffizienten $g_{\text{opt}}(k)$ des *Wiener-Filters* dar. Diese optimalen Filterkoeffizienten ermöglichen die Schätzung mit minimalem Schätzfehler $e_{\min}$ gemäß Gleichung (2.62).

Durch zweiseitige z-Transformation der Gleichung (2.74) für alle u erhält man die Z-Transformierte der *nichtkausalen* Filterimpulsantwort zu:

$$G_{opt}(z) = \frac{R_{xy}(z)}{R_{xx}(z)}, \tag{2.75}$$

wobei $R_{xy}(z)$ das Kreuzleistungsdichtespektrum (KLDS) zwischen Ausgangssignal $y(k)$ und Eingangssignal $x(k)$ und $R_{xx}(z)$ das (Auto-)Leistungsdichtespektrum (LDS) des Eingangssignals des Optimalfilters darstellen. Der Frequenzgang des Filters ergibt sich durch die Substitution $z = e^{j\Omega}$ zu

$$G_{opt}(\Omega) = \frac{R_{xy}(\Omega)}{R_{xx}(\Omega)}. \tag{2.76}$$

2.6.2 Kausales Wiener Filter

Ein nichtkausales Filter nach (2.75) läßt sich technisch nicht realisieren. Es kann aber unter bestimmten Voraussetzungen durch ein kausales FIR-Filter mit endlicher Anzahl von Koeffizienten approximiert werden. Die Berechnung der Koeffizienten des FIR-Filters der Ordnung N ist im allgemeinen nicht durch Frequenztransformation möglich. Die diskrete Wiener-Hopf-Gleichung soll nun für ein kausales Optimalfilter gelöst werden. Dies bedeutet, daß $g_{\text{opt}} = 0$ für $k < 0$ gefordert wird. Das Ausgangssignal des Filters ist durch die endliche Summe

$$\hat{y}(k) = \sum_{u=0}^{N} g(u)x(k-u) = \mathbf{G}^T\mathbf{X}(k) \tag{2.77}$$

gegeben, wobei

$$\mathbf{G} = [g(0), g(1), ..., g(N)]^T \tag{2.78}$$

den Vektor der Filterkoeffizienten und

$$\mathbf{X}(k) = [x(k), x(k-1), ..., x(k-N)]^T \tag{2.79}$$

den Vektor der Abtastwerte des Eingangssignals bezeichnet. Die Minimierung des mittleren quadratischen Fehlers führt unter Anwendung des Orthogonalitätsprinzips aus (2.67) auf ein Gleichungssystem mit $N + 1$ Gleichungen

$$r_{xy}(v) = \sum_{u=0}^{N} g_{\text{opt}}(u)r_{xx}(u-v) \tag{2.80}$$

bzw. in vektorieller Schreibweise

$$\mathbf{p} = \mathbf{R}_{xx}\mathbf{G}_{\text{opt}} \tag{2.81}$$

mit

$$\mathbf{R}_{xx} = E\{\mathbf{X}(k)\mathbf{X}^T(k)\} = \begin{pmatrix} r_{xx}(0) & r_{xx}(1) & \cdots & r_{xx}(N) \\ r_{xx}(1) & r_{xx}(0) & \cdots & r_{xx}(N-1) \\ \cdots & \cdots & \cdots & \cdots \\ r_{xx}(N) & r_{xx}(N-1) & \ldots & r_{xx}(0) \end{pmatrix} \tag{2.82}$$

und

$$\mathbf{p} = E\{y(k)\mathbf{X}(k)\} = [r_{xy}(0), r_{xy}(1), ..., r_{xy}(N)]^T. \tag{2.83}$$

Die optimalen Filterkoeffizienten $\mathbf{G}_{\text{opt}}$ erhält man durch Inversion der Korrelationsmatrix mit:

$$\mathbf{G}_{\text{opt}} = \mathbf{R}_{xx}^{-1}\mathbf{p}. \tag{2.84}$$

Da die Korrelationsmatrix Toeplitzstruktur aufweist, kann das Gleichungssystem in (2.81) effizient z.B. mit dem *Levinson-Durbin-Algorithmus* gelöst werden.

2.6.3 Fehlerbetrachtung

In Abschnitt 2.6.2 wurde das kausale FIR Wiener Filter mit *N* Koeffizienten berechnet. Genau genommen wurde von Wiener und Kolomogoroff ein Filter entwickelt, dessen Gewichtsfunktion nicht auf endliche Dauer begrenzt ist, so daß die Optimalität erst dann erreicht wird, wenn alle Einschwingvorgänge des Filters abgeschlossen sind.

Es stellt sich die Frage, wie sich der endliche Grad des FIR-Filters auf den minimalen Schätzfehler des approximierten Optimalfilters auswirkt. Dazu wird der erreichbare minimale Fehler bei nichtkausalem Filter berechnet:

$$\begin{aligned} E\{e^2(k)\}_{\min} &= E\left\{\left[\, y(k) - \hat{y}_{\text{opt}}(k)\,\right]^2\right\} \\ &= E\left\{y(k)\left[\, y(k) - \hat{y}_{\text{opt}}(k)\,\right] - \hat{y}_{\text{opt}}(k)\left[\, y(k) - \hat{y}_{\text{opt}}(k)\,\right]\right\} \\ &= E\left\{y(k)\cdot e_{\min} - \hat{y}_{\text{opt}}(k)\cdot e_{\min}\right\}. \end{aligned} \tag{2.85}$$

Hierbei verschwindet wegen der Orthogonalitätsbedingung $E\{\hat{y}_{\text{opt}}(k)\cdot e_{\min}\} = 0$ der zweite Summand in (2.85), so daß damit und mit (2.61) folgt:

$$\begin{aligned} E\{e^2(k)\}_{\min} &= E\{y^2(k)\} - E\{y(k)\cdot \sum_{u=-\infty}^{\infty} g_{\text{opt}}(u)x(k-u)\} \\ &= r_{yy}(0) - \sum_{u=-\infty}^{\infty} g_{\text{opt}}(u) r_{xy}(u). \end{aligned} \tag{2.86}$$

Für die folgende Berechnung soll nach Abbildung 2.1 $y(k) = s(k)$ gelten, was bedeutet, daß in diesem Fall $h(i) = 1$ ist. Das Nutzsignal $s(k)$ wird demnach auf dem Übertragungsweg nicht verändert oder verzögert. Dementsprechend wird (2.86) zu:

$$\begin{aligned} E\{e^2(k)\}_{\min} &= E\{s^2(k)\} - E\{s(k)\cdot \sum_{u=-\infty}^{\infty} g_{\text{opt}}(u)x(k-u)\} \\ &= r_{ss}(0) - \sum_{u=-\infty}^{\infty} g_{\text{opt}}(u) r_{xs}(u). \end{aligned} \tag{2.87}$$

Mit der Parseval'schen Beziehung

$$\sum_{n=-\infty}^{\infty} x(n)y(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\Omega)Y^*(\Omega)d\Omega \tag{2.88}$$

und der Substitution von (2.76) gilt dann:

$$\begin{aligned} E\{e^2(k)\}_{\min} &= \frac{2}{\pi}\int_{-\pi}^{\pi} [R_{ss}(\Omega) - \frac{R_{xs}(\Omega)R_{sx}(\Omega)}{R_{xx}(\Omega)}]d\Omega \\ &= \frac{2}{\pi}\int_{-\pi}^{\pi} R_{ss}(\Omega)\cdot[1 - C_{xs}(\Omega)]d\Omega, \end{aligned} \tag{2.89}$$

wobei $C_{xs}(\Omega)$ lt. Gleichung (2.43) die Kohärenz zwischen dem gestörten Eingangssignal $x(k)$ und dem Nutzsignal $s(k)$ darstellt. Unter der Voraussetzung, daß $x(k) = s(k) + n(k)$ und Nutzsignal $s(k)$ und Störung $n(k)$ orthogonal sind, ergibt sich entsprechend für den minimalen mittleren quadratischen Fehler:

$$E\{e^2(k)\}_{\min} = \frac{2}{\pi}\int_{-\pi}^{\pi} R_{ss}(\Omega)\cdot[1-\frac{R_{ss}(\Omega)}{R_{ss}(\Omega)+R_{nn}(\Omega)}]d\Omega. \tag{2.90}$$

Ist die spektrale Leistungsdichte der Störung $R_{nn}(\Omega)$ bezogen auf die spektrale Leistungsdichte des Nutzsignales $R_{ss}(\Omega)$ klein, so ist auch der minimale Fehler klein.

Um nun die Auswirkungen einer begrenzten Filterlänge *N* eines kausalen Filters mit der Beziehung (2.90) zu vergleichen, soll auch für diesen Fall der minimale mittlere quadratische Fehler berechnet werden. Gleichung (2.87) ändert sich für den kausalen Fall und den Filtergrad *N* zu:

$$\begin{aligned} E\{e^2(k)\}_{\min} &= E\{s^2(k)\} - E\{s(k)\cdot\sum_{u=0}^{N} g_{\text{opt}}(u)x(k-u)\} \\ &= r_{ss}(0) - \sum_{u=0}^{N} g_{\text{opt}}(u)r_{xs}(u). \end{aligned} \tag{2.91}$$

Da die Leistung des minimalen Fehlers nicht größer als die Leistung des zu approximierenden Signals sein kann, folgt

$$\sum_{u=0}^{N} g_{\text{opt}}(u)r_{xs}(u) > 0. \tag{2.92}$$

Geht man von $g_{\text{opt}}(u<0) = 0$ aus, so unterscheidet sich der minimale mittlere quadratische Fehler des nichtkausalen Optimalfilters vom kausalen Optimalfilter der Länge *N* nur noch durch:

$$\begin{aligned} \Delta E\{e^2(k)\}_{\min}^{N} &= E_{\min}\{e^2(k)\}_{kausal} - E_{\min}\{e^2(k)\}_{nichtkausal} \\ &= \sum_{u=N}^{\infty} g_{\text{opt}}(u)r_{xs}(u) > 0 \end{aligned} \tag{2.93}$$

und für ein Filter der Länge $L < N$ gilt entsprechend:

$$\begin{aligned}\Delta E\{e^2(k)\}_{\min}^{L} &= \sum_{u=L}^{\infty} g_{\text{opt}}(u) r_{xs}(u) \\ &= \sum_{u=N}^{\infty} g_{\text{opt}}(u) r_{xs}(u) + \sum_{u=L}^{N} g_{\text{opt}}(u) r_{xs}(u) \\ &= \Delta E\{e^2(k)\}_{\min}^{N} + \sum_{u=L}^{N} g_{\text{opt}}(u) r_{xs}(u) > 0.\end{aligned} \tag{2.94}$$

Durch die Verringerung des Filtergrades um eins ergibt sich demnach der daraus zusätzlich entstehende Fehler zu:

$$\Delta E\{e^2(k)\}_{\min}^{N-L} - \Delta E\{e^2(k)\}_{\min}^{N} = \Delta E\{e^2(k)\}_{\min}^{L} = \sum_{u=L}^{N} g_{\text{opt}}(u) r_{xs}(u) > 0. \tag{2.95}$$

Durch vollständige Induktion kann leicht gezeigt werden, daß die Erhöhung der Filterordnung den minimalen mittleren quadratischen Fehler verringert und der minimale mittlere quadratische Fehler des nichtkausalen Filters eine untere Schranke für das kausale Optimalfilter darstellt. Eine Erhöhung der Filterordnung des optimalen Filters ist gemäß (2.95) nur dann sinnvoll, wenn die Kreuzkorrelationsfunktion $r_{xs}(u)$ im Bereich $L \le u \le N$ nicht verschwindend kleine Werte besitzt.

2.7 Spektrale Subtraktion

Ausgehend von Abbildung 2.1 soll ein allgemeines vorerst nichtkausales Wiener Filter implementiert werden. Das Filter soll aus dem Signal

$$x(k) = s(k) + n(k) \tag{2.96}$$

das Nutzsignal $s(k)$ und das Störsignal $n(k)$ möglichst gut separieren.

Als Eingangssignal steht nur das Signalgemisch $x(k)$ gemäß Gleichung (2.96) zur Verfügung. Alle drei Signale sind stationär, die Kreuzkorrelationsfunktionen und Leistungsdichtespektren der Signale existieren. Über die Signalstatistik sei a priori nichts bekannt. Die gewünschte Ausgangsfunktion des unbekannten Systems $y(k)$ bestimmt sich nach Abbildung 2.1 zu:

$$y(k) = \sum_{v=-\infty}^{\infty} h(v) s(k-v). \tag{2.97}$$

Das Filter ist so zu optimieren, daß der mittlere quadratische Fehler minimal wird. Unter Ausnutzung des Orthogonalitätsprinzips lt. (2.67) und durch Verwenden von (2.61) und (2.62) findet man mit dem Ansatz (2.73):

$$\sum_{v=-\infty}^{\infty} h(v) r_{xx}(u-v) - \sum_{v=-\infty}^{\infty} h(v) r_{xn}(u-v) - \sum_{v=-\infty}^{\infty} g_{\text{opt}}(v) r_{xx}(u-v) = 0. \tag{2.98}$$

Nach z-Transformation ergibt sich für die optimalen Koeffizienten des nichtkausalen Wiener-Filters:

$$G_{\text{opt}}(z) = H(z) \cdot \frac{R_{xx}(z) - R_{xn}(z)}{R_{xx}(z)} = H(z) \cdot \left. \frac{R_{xx}(z) - R_{nn}(z)}{R_{xx}(z)} \right|_{s \perp n}. \tag{2.99}$$

Die Übertragungsfunktion des allgemeinen, nichtkausalen Wiener Filters bestimmt sich demnach durch die lineare Übertragungsfunktion $H(z)$, die durch einen frequenzabhängigen Faktor korrigiert wird. Dieser Faktor ist abhängig von der Leistungsdichte des gestörten Signals $x(k)$ und der Kreuzleistungsdichte zwischen Störung $n(k)$ und Signalgemisch $x(k)$. Entsprechend der Subtraktion in der Darstellung (2.99) wird diese Art der Geräuschreduktion durch ein optimales Wiener Filter mit *Spektraler Subtraktion* bezeichnet.

Ein Sonderfall in Gleichung (2.99) ergibt sich, wenn Nutzsignal und Störung orthogonal zueinander sind. Eine weit verbreitete Darstellung der Übertragungsfunktion des Wiener Filters gibt

$$\begin{aligned} G_{\text{opt}}(z) &= H(z) \cdot \frac{R_{ss}(z)}{R_{ss}(z) + R_{nn}(z)} \\ &= H(z) \cdot \left[1 - \frac{R_{nn}(z)}{R_{xx}(z)} \right] \end{aligned} \tag{2.100}$$

wieder. Nach Substitution von $z = e^{j\Omega}$ ergibt sich für den Frequenzgang des optimalen Wiener Filters:

$$G_{\text{opt}}(\Omega) = H(\Omega) \cdot \left[1 - \frac{R_{nn}(\Omega)}{R_{xx}(\Omega)} \right]. \tag{2.101}$$

In analoger Weise lassen sich die Filterkoeffizienten des *kausalen* Wiener Filters *N*-ten Grades bestimmen. Mit

$$\mathbf{H} = [h(0), h(1), ..., h(N)]^T, \tag{2.102}$$

$$\mathbf{q}_x = E\{x(k)\mathbf{X}(k)\} = [r_{xx}(0), r_{xx}(1), ..., r_{xx}(N)]^T \tag{2.103}$$

und

$$\mathbf{q}_n = E\{n(k)\mathbf{X}(k)\} = [r_{xn}(0), r_{xn}(1), ..., r_{xn}(N)]^T \tag{2.104}$$

folgt dann letztendlich für die N Koeffizienten des kausalen Wiener Filters:

$$\mathbf{G}_{\text{opt}} = \mathbf{R}_{xx}^{-1}\mathbf{H}^T(\mathbf{q}_x - \mathbf{q}_n). \tag{2.105}$$

In (2.99) taucht die Kreuzleistungsdichte $R_{xn}(z)$ auf, die explizit nicht zur Verfügung steht. Auch die Störleistungsdichte $R_{nn}(\Omega)$ und Leistungsdichte des ungestörten Nutzsignals $R_{ss}(\Omega)$ in Beziehung (2.100) liegen nicht vor. Diese müssen entsprechend geschätzt werden. Hierfür haben sich in der Literatur verschiedene Schätzverfahren etabliert, auf die im Vergleich zu einem neu entwickelten Verfahren im Abschnitt 6.2 eingegangen wird.

2.8 Ephraim-Malah-Filter

Das Wiener Filter minimiert den mittleren quadratischen Fehler zwischen ungestörtem Nutzsignal und der Schätzung des Nutzsignals. Das menschliche Gehör reagiert auf Phasenstörungen relativ unempfindlich, während Änderungen der Amplitude sehr gut wahrgenommen werden, siehe Abschnitt 3.2.2.

Aus diesem Grund ist es sinnvoll, den mittleren quadratischen Fehler zwischen dem Betragsspektrum des ungestörten Nutzsignals und dem Betragsspektrum des geschätzten Signals zu minimieren. Bei dieser Modifikation des Optimierungsproblems wird das Phasenspektrum nicht verändert. Dieses in [49] vorgeschlagene Verfahren führt auf das sogenannte Ephraim-Malah-Filter. Das komplexe Spektrum des gestörten Signals $x(k)$ wird mit $X(\Omega)$ bezeichnet und das Spektrum des ungestörten Signals $s(k)$ wird durch

$$S(\Omega) = A_s(\Omega) e^{j\Phi_s(\Omega)} \tag{2.106}$$

dargestellt, wobei Betrag $A_s(\Omega)$ und Phase $\Phi_s(\Omega)$ getrennt sind. Die MMSE-Schätzung des Betragsspektrums $A_s(\Omega)$ des Sprachsignals ergibt sich aus folgendem Ansatz:

$$\hat{A}_s(\Omega) = E\left\{W\left[A_s(\Omega)\middle|X(\Omega)\right]\right\}, \tag{2.107}$$

der aus der a-posteriori-Wahrscheinlichkeitsverteilung $W[A_s(\Omega)|X(\Omega)]$ resultiert. Die Lösung nach [49], umgeformt nach der Übertragungsfunktion des Optimalfilters lautet:

$$G_{opt}(\Omega) = W\{H_s\,|X(\Omega)\} \cdot \frac{\sqrt{\pi}}{2} \frac{\sqrt{\upsilon(\Omega)}}{\psi(\Omega)} \Upsilon(-0.5;1;-\upsilon(\Omega)) \tag{2.108}$$

mit

$$\upsilon(\Omega) = \frac{\xi(\Omega)}{1+\xi(\Omega)} \psi(\Omega), \tag{2.109}$$

wobei das a-priori-SNR

$$\xi(\Omega) = \frac{R_{ss}(\Omega)}{R_{nn}(\Omega)} \tag{2.110}$$

und das a-posteriori-SNR

$$\psi(\Omega) = \frac{R_{xx}(\Omega)}{R_{nn}(\Omega)} \tag{2.111}$$

beträgt und

$$\Upsilon(-0.5;1;-\upsilon(\Omega)) = \left[(1+\upsilon(\Omega)) I_0\left(\frac{\upsilon(\Omega)}{2}\right) + \upsilon(\Omega) I_1\left(\frac{\upsilon(\Omega)}{2}\right)\right] e^{-\frac{\upsilon(\Omega)}{2}} \tag{2.112}$$

eine konfluent hypergeometrische Funktion mit den modifizierten Besselfunktionen $I_{0,1}(.)$ nullter und erster Ordnung bezeichnet.

In (2.109) tritt der Term $W\{H_s|X(\Omega)\}$ auf. Dazu wird in [49] ein Wahrscheinlichkeitsmodell beschrieben, mit dem festgestellt wird, ob im Spektrum $X(\Omega)$ des gestörten Nutzsignals $x(k)$ auch wirklich Komponenten des Sprechersignals $s(k)$ auftreten, also in welchen spektralen Bereichen Sprecheraktivität vorliegt. Danach ergibt sich

$$W\{H_s \mid X(\Omega)\} = \frac{1}{1+\frac{W_p}{W_s}(1+\xi(\Omega))e^{-\upsilon(\Omega)}}, \tag{2.113}$$

wobei H_s für die Hypothese „Sprecher spricht“ steht und W_p sowie W_s die Wahrscheinlichkeiten für Sprachpause bzw. Sprachaktivität bezeichnen.

3 Spracherzeugung und -perzeption

Die Nutzung psychoakustischer Effekte für die latente Systemidentifikation und Geräuschreduktion erfordert das Verständnis der menschlichen Sprachproduktion und -perzeption. Die Modellierung des Spracherzeugungsprozesses und die Analyse des Sprachsignals ergeben wichtige Merkmale, die für die spätere Sprachsignalverbesserung, Spracherkennung und Synthese große Bedeutung haben. Das menschliche Gehör ist der Empfänger der akustischen Information. Aus seinem Aufbau und der Funktionsweise ergeben sich Modelle der akustischen Wahrnehmung und Effekte, die für die auditive Signalverbesserung anhand eines Geräuschreduktionssystems oder für die psychoakustische Identifikation des Übertragungskanals, beispielsweise in einer Echokompensation, genutzt werden können.

3.1 Das menschliche Sprechorgan

Die Bildung des Sprachsignals erfolgt prinzipiell in zwei Stufen: *Anregung* und *Signalformung*. Unter *Anregung* versteht man den Vorgang, bei dem durch Variation des Luftstromes Schwingungen oder Geräusche erzeugt werden, die nach weiterer *Formung* im Sprechtrakt als Schall wahrnehmbar sind. Die Lungen stellen dabei den notwendigen Luftstrom und die Anregungsenergie zur Verfügung. Es werden drei Anregungsarten unterschieden: Die Glottis formt den Luftstrom durch Gegeneinanderschwingen der beiden Stimmbänder und erzeugt so *stimmhafte* Anregung (Phonation). Bei Phonation erzeugen die Stimmbänder eine näherungsweise periodische, nichtsinusförmige Schwingung. Durch die Vorspannung der Stimmbänder (abhängig vom Alter und Geschlecht) und die Stärke des subglottalen Luftdrucks ergibt sich die Amplitude sowie die Grundfrequenz f_0 der einzelnen Schwingungszyklen der Glottis. Bleiben die Stimmbänder geöffnet, durchströmt die Luft eine Engstelle im Mund- oder Rachenraum und erzeugt dort eine rauschartige turbulente Strömung. Diese bildet das Anregungssignal bei *stimmloser* Anregung (Frikation). Durch plötzliches Öffnen eines supraglottalen Verschlusses entsteht *transiente* Anregung (Plosivanregung), die durch ihren Zeitverlauf charakterisiert ist. Zunächst wird durch den Verschluß im Mund- oder Rachenraum nahezu jeder Schall unterbunden, es entsteht eine kurze Pause. Das Öffnen des Verschlusses bewirkt das Entweichen der angestauten Luft mit einem spezifischen Plosionsgeräusch.

Das Anregungssignal reicht für die Bildung der Sprache nicht aus. Es bedarf noch der *Signalformung* in dem aus Rachen- und Mundraum gebildeten *Vokaltrakt*. Der deutlich hörbare Unterschied zwischen nichtnasalen und nasalen Lauten wird durch die Verlagerung des Velums (Gaumensegel) herbeigeführt. Liegt das Velum nicht an der hinteren Rachenwand an, so ist der Nasenraum mit der Mundhöhle und dem Rachen verbunden. Abbildung 3.1 zeigt den menschlichen Sprechtrakt schematisch im Schnitt.

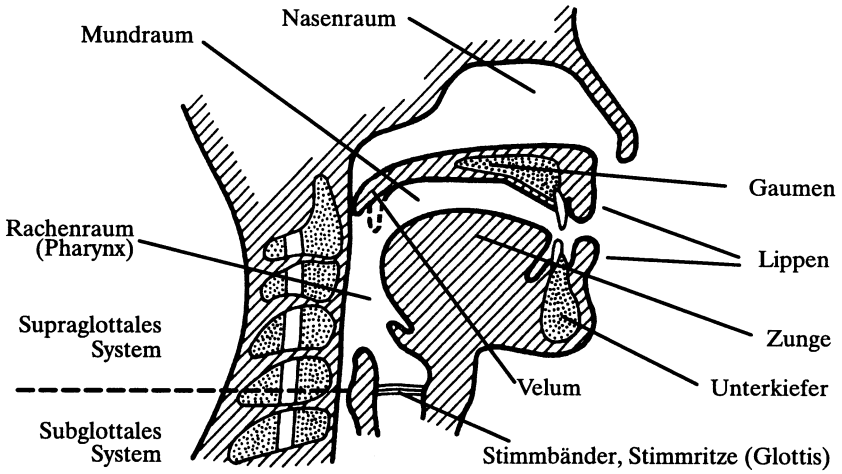


Abbildung 3.1 Schematische Übersicht des Sprechorgans aus [183]. Bestandteile des Sprechtraktes sind die Lungen als Lieferant des Luftstromes, Luftröhre und Bronchien, Kehlkopf und Stimmbänder, die Begrenzung von Rachen- und Mundraum, Velum, Gaumen, Zunge und Lippen und der Nasenraum.

3.1.1 Modellierung des Spracherzeugungsprozesses

Für die Analyse des Sprachsignals und die analytische und methodische Behandlung sprachsignalverarbeitender Verfahren hat die folgende Modellierung des Spracherzeugungsprozesses große Bedeutung. Ausführliche Modelle der einzelnen Sprachkomponenten und des Spracherzeugungsprozesses sind in [43], [51] und [183] zu finden.

Sprachsignale sind im allgemeinen zeitvariant und können nur in kurzen Zeitabständen, üblicherweise innerhalb von 10 ... 20 ms, als quasi stationär betrachtet werden. Außerdem sind sie bei $f_{\max}$ bandbegrenzt. Primär entsteht aus dem Sprachschall durch Wandlung, z.B. in einem Mikrophon, ein zeitkontinuierliches elektrisches Signal $s(t)$. Durch die Abtastung mit der Abtastfrequenz $f_a \geq 2f_{\max}$ ergibt sich ein vollständig rekonstruierbares zeitdiskretes Sprachsignal $s(k)$. Wegen der abschnittsweisen Stationarität der Sprache lassen sich die aus Abschnitt 2.2 bekannten Verfahren der DFT und z-Transformation anwenden. Da üblicherweise der Zeitpunkt k so gewählt ist, daß $s(k<0) = 0$, eignen sich für die Transformation des Sprachsignals die einseitigen z-, Fourier-, und Laplace-Transformationen.

3.1.1.1 Das Fant'sche Source-Filter Modell

In Abbildung 3.2 ist das Gesamtmodell der Spracherzeugung nach Fant [52] dargestellt. Alle Signale und Systeme werden als zeitdiskret betrachtet und in Form ihrer z-Transformierten dargestellt. Stimmhafte und stimmlose Anregung werden auf getrennten Zweigen realisiert. Bei stimmhafter Anregung gibt der Impulsgenerator $U_s(z)$ eine Impulsfolge mit der zeitvarianten Grundperiodendauer T_0 ab. Das Formfilter $G(z)$ erzeugt daraus eine Stimmbandschwingung $U(z)$. Die Stimmbandschwingung regt den Vokaltrakt mit der zeitveränderlichen Übertragungsfunktion $H(z)$ an.

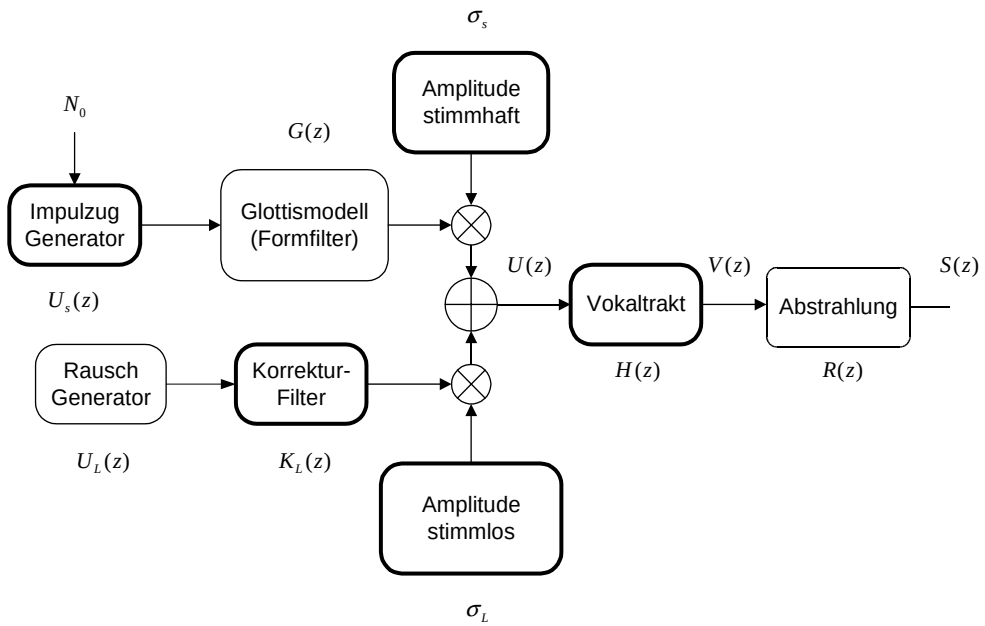


Abbildung 3.2 Fant'sches Source-Filter Modell. Modifikation von [52] und [183]. Das Additionsglied ermöglicht gemischte Anregung. Die Übertragungsfunktionen und Signale sind in zeitdiskreter, zeitinvarianter Darstellung im z-Bereich angegeben. Dick umrandete Teilsysteme markieren zeitvariante Parameter und Komponenten.

Im stimmlosen Zweig erfolgt die Anregung durch einen Rauschgenerator $U_L(z)$, dessen Ausgangssignal anschließend das Korrekturfilter $K_L(z)$ durchläuft. Durch das Korrekturfilter $K_L(z)$ wird die Tatsache berücksichtigt, daß die Übertragungsfunktion des Vokaltrakts $H(z)$ bei stimmhafter und stimmloser Anregung nicht identisch ist. Das gesamte System ist zeitvariant. In kurzen Zeitfenstern von ca. 10 bis 20 ms Dauer kann das System als quasi zeitinvariant angesehen werden. Mit der Einführung des Fensterindex m ergibt sich für die Erzeugung des Sprachsignals $s(k)$ folgende Kurzzeitdarstellung im z-Bereich:

$$S(z,m) = U(z,m)H(z,m)R(z). \tag{3.1}$$

3.1.1.2 Anregungsmodell

Während der *stimmhaften* Anregung (Phonation) strömt die Luft aus den Lungen durch die Glottis und wird dabei moduliert. Der Impulszuggenerator $U_s(z)$ gibt eine Impulsfolge mit der zeitveränderlichen Grundperiodendauer T_0 ab. Das Formfilter $G(z)$ erzeugt daraus die Stimmbandschwingung, die durch eine Dreieckschwingung mit der Grundfrequenz $f_0 = 1/T_0$ approximiert werden kann. Damit ist $G(z)$ in erster Näherung ein Tiefpaßfilter, dessen Übertragungsfunktion sprecherspezifisch ist und mit ca. 12 dB/Oktave abfällt [183]. Durchströmt die Luft bei geöffneten Stimmbändern einen Engpaß im Mund- oder Rachenraum, so entsteht eine rauschartige turbulente Strömung. Der Rauschgenerator $U_L(z)$ bildet so das Anregungssignal für *stimmlose* Laute. σ_S und σ_L sind die Verstärkungsfaktoren, wie in Abbildung 3.2 dargestellt. Sie bestimmen die Amplituden für die stimmhafte bzw. stimmlose Anregung. Für den stimmhaften Fall gilt somit:

$$U(z,m) = \sigma_s U_s(z,m) G(z) \tag{3.2}$$

und für den stimmlosen Fall ergibt sich die Vokaltraktanregung $U(z, m)$ zu:

$$U(z,m) = \sigma_L U_L(z) K_L(z,m). \tag{3.3}$$

Transiente Anregung liegt dann vor, wenn irgendwo im supraglottalen System der Luftstrom kurzzeitig durch einen Verschluß unterbrochen wird und sich dann durch plötzliches Öffnen des Verschlusses die Luft plosionsartig ausbreitet. Die transiente Anregung hat damit ein charakteristisches Zeitverhalten. Die kurze Strömungspause hat eine Dauer von ca. 20 bis 100 ms. Das Plosionsgeräusch dauert ca. 20 bis 50 ms, siehe [183].

3.1.1.3 Vokaltraktmodell

Schallwellen sind Longitudinalwellen. Die Teilchen bewegen sich in Richtung der Ausbreitung des Schalls; Schallschnelle v und Ortskoordinate x sind gleichgerichtet. Grundsätzlich kann sich eine Welle nur dann ungestört ausbreiten, wenn die Schallfeldimpedanz Z gleich bleibt. In diesem Fall erfolgt die Ausbreitung in Form einer Kugelwelle. In einem zylindrischen Rohr, kann sich die Welle dagegen nur noch in der Längsrichtung des Rohres ausbreiten.

Die Schallfeldimpedanz ist folgendermaßen definiert:

$$Z = \frac{\rho_0 c}{A}, \tag{3.4}$$

wobei c die Schallausbreitungsgeschwindigkeit, ρ_0 die Luftdichte bei normalem Luftdruck und A den Querschnitt des Rohres bezeichnen. Das Röhrenmodell des Vokaltrakts stellt den Bezug zu einer Allpol-Übertragungsfunktion her. Dabei wird nach Fant [51] der Vokaltrakt als akustisches Ansatzrohr der Länge L angesehen. Dieses setzt sich aus $i = 0, \ldots, M$ gleich langen Zylinderabschnitten mit variierendem Querschnitt A_i zusammen, siehe Abbildung 3.3. Das Rohr ist am einen Ende offen (Mundseite) und am anderen Ende (Glottis) geschlossen. Da die Wellenlänge des akustischen Signals sehr viel größer als die Länge eines Zylinderabschnitts ist, breitet sich eine ebene Welle in Achsrichtung aus.

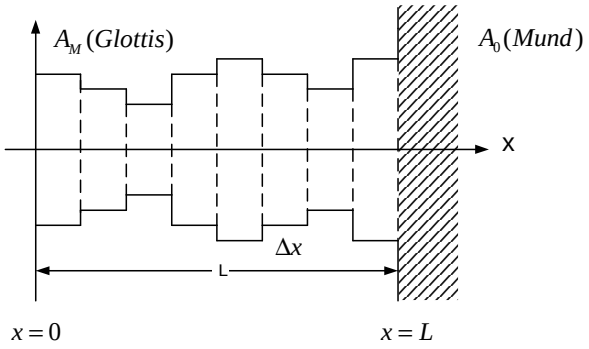


Abbildung 3.3 Röhrenmodell des Vokaltrakts: zylindrisches Ansatzrohr mit unstetigem Querschnittsverlauf und Abschluß durch schallharte Wand (Glottis) und durch freies Schallfeld (Außenwelt). Die Indizierung der Querschnitte A_i erfolgt mit Hinblick auf die digitale Modellierung entgegen der Ausbreitungsrichtung der vorlaufenden Schallwelle.

Ändert sich die Schallfeldimpedanz Z sprunghaft, so kommt es zu Reflexionen. Eine ideal schallharte ebene Wand senkrecht zur Ausbreitungsrichtung der Schallwelle bewirkt beispielsweise eine vollständige Reflexion der Welle. Da sich die Teilchen an der schallharten Wand nicht mehr bewegen lautet die Randbedingung für die Reflexion $v = 0$.

Auch beim Auftreffen der Welle auf ein ideal schallweiches Medium treten Reflexionen auf. Endet beispielsweise ein zylindrisches Rohr, in dem sich eine ebene Schallwelle ausbreitet, mit einer Öffnung ins freie Schallfeld, so kann sich die Welle ab dort frei in alle Richtungen ausbreiten. Dieser Übergang hat die Randbedingung $p = 0$. Durch die Überlagerung von ursprünglicher und reflektierter Welle entsteht eine stehende Welle. Nach *Ungeheuer* [178] kann das Ansatzrohr aus Abbildung 3.3 mit Hilfe der Webster'schen Horngleichung modelliert werden. Demnach wird die Ausbreitung einer ebenen Welle durch ein verlustfreies akustisches Rohr mit ortsveränderlichem Querschnitt $A(x)$ beschrieben. Es gilt:

$$\frac{\partial^2 \phi(x,t)}{\partial x^2} + \frac{1}{A(x)} \frac{dA}{dx} \frac{\partial \phi(x,t)}{\partial x} = \frac{1}{c^2} \frac{\partial^2 \phi(x,t)}{\partial t^2}. \tag{3.5}$$

Hierbei stellt $\phi(x, t)$ das *Geschwindigkeitspotential* dar, eine Hilfsgröße, die folgendermaßen definiert ist:

$$p = \rho_0 \frac{\partial \phi(x,t)}{\partial t} \text{ und } v = -\frac{\partial \phi(x,t)}{\partial x}. \tag{3.6}$$

Das Geschwindigkeitspotential $\phi(x, t)$ wird eingeführt, damit die Randbedingungen an Mund und Glottis in der gleichen Variablen ausgedrückt werden können.

In der Regel ist Gleichung (3.5) in geschlossener Form nicht lösbar. Für den Fall eines einfachen Rohrabschnitts mit konstantem Querschnitt A_i wird Gleichung (3.5) für den verlustfreien Fall zur gewöhnlichen Wellengleichung:

$$\frac{\partial^2 \phi(x,t)}{\partial x^2} - \frac{1}{c^2} \frac{\partial^2 \phi(x,t)}{\partial t^2} = 0. \tag{3.7}$$

Nach *d'Alembert* lassen sich dafür allgemeine Lösungen der Form

$$\phi(x,t) = \phi_f \left(t - \frac{x}{c} \right) + \phi_b \left(t + \frac{x}{c} \right) \tag{3.8}$$

angeben. Dabei bezeichnen c die Schallausbreitungsgeschwindigkeit, ρ_0 die Dichte der Luft, ϕ_f die forwärtslaufende und ϕ_b die zurücklaufende Welle.

Mit der Einführung des *Volumengeschwindigkeit* u mit $u = v \cdot A$, wobei u_f, p_f die jeweils hinlaufenden Wellen und u_b, p_b die jeweils zurücklaufenden Wellen bezeichnen, folgen für die Volumengeschwindigkeit u und den Schalldruck p:

$$\begin{aligned} u(x,t) &= \frac{A}{c} \left[\phi_f \left(t - \frac{x}{c} \right) + \phi_b \left(t + \frac{x}{c} \right) \right] \\ &= u_f \left(t - \frac{x}{c} \right) + u_b \left(t + \frac{x}{c} \right) \end{aligned} \tag{3.9}$$

und:

$$\begin{aligned} p(x,t) &= \frac{A}{c}\rho_0\left[\phi_f\left(t-\frac{x}{c}\right)+\phi_b\left(t+\frac{x}{c}\right)\right] \\ &= p_f\left(t-\frac{x}{c}\right)+p_b\left(t+\frac{x}{c}\right). \end{aligned} \tag{3.10}$$

Unter Berücksichtigung der akustischen Impedanz Z wird der Druck p auf die Volumengeschwindigkeit u zurückgeführt. Mit der Definition der akustischen Impedanz lt. Formel (3.4) ergibt sich aus Gleichung (3.10):

$$p(x,t) = Z\cdot\left[u_f\left(t-\frac{x}{c}\right)+u_b\left(t+\frac{x}{c}\right)\right]. \tag{3.11}$$

Im allgemeinen Fall ist das Modell des Vokaltrakts aus mehreren jeweils homogenen Rohrabschnitten A_i zusammengesetzt. An den unstetigen Stoßstellen können sich jedoch Druck und Volumengeschwindigkeit nicht sprunghaft ändern, d.h. auf beiden Seiten der *i*-ten Stoßstelle (x_i, x_{i-1}) liegen gleicher Druck und Volumengeschwindigkeit vor. Es gilt damit:

$$\begin{aligned} p_i(t) &= p_{i-1}(t) \\ u_i(t) &= u_{i-1}(t). \end{aligned} \tag{3.12}$$

Sind A_i, A_{i-1} die jeweiligen Querschnitte benachbarter Rohrabschnitte und Z_i, Z_{i-1} die dazugehörigen Schallfeldimpedanzen, so gilt an den Abschnittsübergängen für den Schalldruck $p_i(t)$:

$$Z_i[u_{f,i}(t)+u_{b,i}(t)] = Z_{i-1}[u_{f,i-1}(t)+u_{b,i-1}(t)] \tag{3.13}$$

sowie für die Volumengeschwindigkeit $u_i(t)$:

$$u_{f,i}(t)-u_{b,i}(t) = u_{f,i-1}(t)-u_{b,i-1}(t). \tag{3.14}$$

Unter Berücksichtigung von Gleichung (3.4), (3.10) und mit (3.12) beträgt das Verhältnis von Querschnitt und Schallfeldimpedanz an einem Querschnittsübergang:

$$\frac{Z_i}{Z_{i-1}} = \frac{A_{i-1}}{A_i}. \tag{3.15}$$

Nach Definition des Reflexionskoeffizienten r_i zwischen i-tem und $i-1$-tem Querschnitt durch

$$r_i = \frac{A_{i-1} - A_i}{A_{i-1} + A_i} \tag{3.16}$$

erhält man für die vorlaufende Welle im $i-1$-tem Rohrabschnitt:

$$u_{f,i-1}(t) = (1+r_i) \cdot u_{f,i}(t) + r_i \cdot u_{b,i-1}(t) \tag{3.17}$$

und für die rücklaufende Welle im Abschnitt i:

$$u_{b,i}(t) = -r_i \cdot u_{f,i}(t) + (1-r_i) \cdot u_{b,i-1}(t). \tag{3.18}$$

Für jedes Teilstück Δx ergibt sich eine Wellenlaufzeit τ mit

$$\tau = \frac{L}{cM} = \frac{\Delta x}{c}. \tag{3.19}$$

Wird der Verlauf von hinlaufender Welle $u_f(x)$ und rücklaufender Welle $u_b(x)$ nach Reihenschaltung der Teilsysteme mit $t = i \cdot 2\tau = i \cdot T_a$ abgetastet, läßt sich dieses im Zeitbereich analoge Modell in ein Digitalfilter überführen. Für den verlustfreien Fall gilt mit Gleichung (3.17) und Beziehung (3.18):

$$u_f(i-1) = (1+r_i) \cdot u_f(i) + r_i \cdot u_b(i-1) \tag{3.20}$$

sowie:

$$u_b(i) = -r_i \cdot u_f(i) + (1-r_i) \cdot u_b(i-1). \tag{3.21}$$

Verzögerungen um τ entsprechen in der z-Ebene einer Multiplikation mit z^{-1}. Für die Beschreibung des Signalflusses an den Querschnittsübergängen A_{i-1}, A_i und A_{i+1} ergibt sich ein digitales Signalflußmodell für die hin- und rücklaufende Schallwelle, siehe Abbildung 3.4.

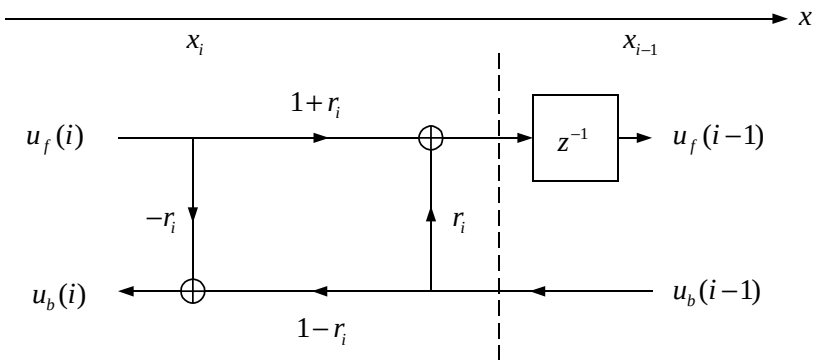


Abbildung 3.4 Zeitdiskretes Signalflußmodell in einem Abschnitt des Vokaltrakts und für den idealen verlustfreien Fall. Im realen Vokaltrakt treten Verluste durch Reibung in der Luft, Wärmeeffekte und Mitschwingen der Vokaltraktwände auf. Der Index i des Vokaltraktabschnitts läuft mit Rücksicht auf die Darstellung von Digitalfiltern in der direkten Form entgegen der Ausbreitungsrichtung des Schalls.

Eine derartige Digitalfilterstruktur, die aus der Theorie der linearen Prädiktion bekannt ist, wird als Viermultiplizierer-Gitterstruktur (*ladder structure*) bezeichnet. Aus der Topologie ist unmittelbar ersichtlich, daß es sich hierbei um ein rein rekursives Digitalfilter handelt. Nach z-Transformation von Gleichung (3.20) und (3.21) und einfacher Umformung erhält man die Beschreibung des Signalflußmodells eines Vokaltraktabschnitts in Matrizenform zu:

$$\begin{bmatrix} U_{f,i}(z) \\ U_{b,i}(z) \end{bmatrix} = \begin{bmatrix} \frac{1}{1+r_i} \cdot z & \frac{-r_i}{1+r_i} \cdot z \\ \frac{-r_i}{1+r_i} & \frac{1}{1+r_i} \end{bmatrix} \cdot \begin{bmatrix} U_{f,i-1}(z) \\ U_{b,i-1}(z) \end{bmatrix} \tag{3.22}$$

und in entsprechend kompakter Form mit:

$$\mathbf{U}_i(z) = \mathbf{\Lambda}_i \cdot \mathbf{U}_{i-1}(z). \tag{3.23}$$

Durch Reihenschaltung mehrerer Vokaltraktabschnitte $\mathbf{U}_i(z)$ mit dem Index $i = 0, 1, \ldots, M$ nach Abbildung 3.3 folgt mit Gleichung (3.23):

$$\mathbf{U}_M(z) = \mathbf{\Lambda}_M \cdot \mathbf{\Lambda}_{M-1} \cdot \ldots \cdot \mathbf{\Lambda}_1 \cdot \mathbf{U}_0(z) = \left[\prod_{i=M}^{1} \mathbf{\Lambda}_i \right] \cdot \mathbf{U}_0(z). \tag{3.24}$$

Das Eingangssignal des Vokaltrakts nach Abbildung 3.2 ist im z-Bereich das Anregungssignal $U(z)$ und das Ausgangssignal des Vokaltrakts ist das Sprachsignal $S(z)$. Mit Gleichung (3.24) und

$$\begin{aligned} \mathbf{U}_0(z) &= \begin{bmatrix} S(z) \\ 0 \end{bmatrix} \\ \mathbf{U}_M(z) &= \begin{bmatrix} U(z) \\ 0 \end{bmatrix} \end{aligned} \tag{3.25}$$

kann die Vokaltraktübertragungsfunktion $H(z)$ wie folgt bestimmt werden. Es gilt nämlich:

$$H(z) = \frac{S(z)}{U(z)} = \frac{1}{\prod_{i=1}^{M} \mathbf{\Lambda}_i^{-1} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix}}. \tag{3.26}$$

Nach Invertierung der Matrizen Λ_i aus Gleichung (3.22) läßt sich die Vokaltraktübertragungsfunktion $H(z)$ aus Beziehung (3.26) in folgende Form bringen:

$$H(z) = \frac{z^{-M/2} \cdot \prod_{i=1}^{M} (1 + r_i)}{\prod_{i=1}^{M} (1 - p_i z^{-1})} = \frac{H_0}{1 - \sum_{i=1}^{M} b_i z^{-i}}. \tag{3.27}$$

Die berechnete Übertragungsfunktion $H(z)$ des Vokaltrakt-Modells hat nur triviale Nullstellen. Die Koeffizienten p_i spiegeln die Paare konjugiert komplexer Polstellen, die sogenannten Vokaltraktresonanzen (*Formanten*) wider. Damit ist die eingangs erwähnte Allpol-Übertragungsfunktion ermittelt worden. Die Vokaltraktübertragungsfunktion $H(z)$ erinnert an die Form autoregressiver Prozesse in zeitdiskreter Darstellung, wobei die Koeffizienten b_i die Multiplikatoren im rekursiven Zweig des entsprechenden Digitalfiltermodells bezeichnen. Ein derartiges Filter wird in Abbildung 3.5 gezeigt.

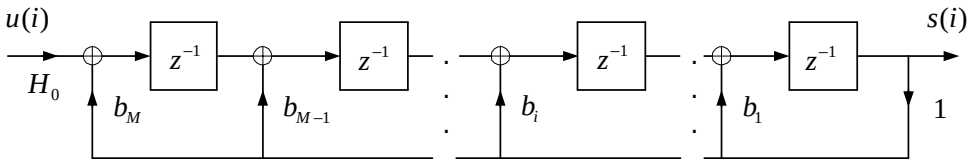


Abbildung 3.5 Digitales Filtermodell des Vokaltrakts in direkter Form laut (3.27)

Bei Ankopplung des Nasenraums kommt es in der kombinierten Vokal-Nasaltrakt-Übertragungsfunktion zu Nullstellen, die sich in der Dämpfung einzelner Frequenzkomponenten (*Antiformanten*) bemerkbar machen. Der Vokaltrakt hat entscheidenden Einfluß auf das resultierende Sprachsignal. Wie sich in späteren Betrachtungen zeigt, hat die Form der Vokaltrakt-Übertragungsfunktion große Ähnlichkeit mit autoregressiven Prozessen. Dadurch wird die parametrische Modellierung des Sprachsignals für die Sprachverbesserung, Geräuschreduktion, Spracherkennung oder z. B. Sprachsynthese erleichtert. Vor allem in der Codier- und Komprimiertechnologie werden diese Spracherzeugungsmodelle angewendet. In den neu entwickelten und in Abschnitt 6 vorgestellten Verfahren, werden Modelle der Spracherzeugung für die LPC- und Kepstral-Analysen des gestörten Eingangssignals genutzt, um daraus später Anhaltspunkte für die Bestimmung einer psychoakustischen Geräuschreduktion zu gewinnen.

3.2 Das menschliche Gehör

Das Außen-, Mittel- und Innenohr bilden das Hörorgan, siehe Abbildung 3.6. Der Schall wird durch das Außenohr aufgefangen und im ca. 20 mm langen Gehörgang zum Trommelfell weitergeleitet. Während sich im Außenohr der Schall noch in der Luft ausbreitet, erfolgt im Innenohr die Leitung des Schalls in Lymphflüssigkeiten. Aufgrund der Unterschiede in der akustischen Ausbreitung der Schallwellen übernimmt das Mittelohr die Anpassung vom Außenohr zum Innenohr. Dies geschieht durch mechanische Wandlung über die Gehörknöchelchen. Dabei werden die vom Trommelfell aufgenommenen Schwingungen zum ovalen Fenster, dem Eingang zum Innenohr, geleitet. Die drei Knöchelchen sind beweglich gelagert und wirken wie kleine Hebel. Neben der Hebelübersetzung wirkt auch die Flächenübersetzung vom relativ großen Trommelfell zum kleinen ovalen Fenster. Auf diese Weise erfolgt eine zusätzliche Verstärkung des eintreffenden Schalls. Das Innenohr ist in das Felsenbein, einem harten Knochen, eingelagert und mit inkompressibler Lymphflüssigkeit gefüllt. Es hat die Form einer Schnecke (Cochlea) mit ca. 2,5 Windungen. Die Cochlea besteht aus drei parallel verlaufenden Kanälen. Der obere und untere Kanal werden von der Basilarmembran getrennt. Auf dieser Membran liegt das Cortische Organ mit den Sinneszellen des Gehörs.

Die Basilarmembran schwingt in Wanderwellen. Es gibt also keine Schwingungsbäuche oder Knoten. Vielmehr wandert die Welle auf der Basilarmembran in Richtung Helicotrema, einer Verbindung von oberen und unterem Kanal. Auf diese Weise entsteht ein für den Hörvorgang äußerst wichtiger Effekt, nämlich die *Frequenz-Orts-Transformation* auf der Basilarmembran, siehe Abbildung 3.7. In Abschnitt 3.3 werden psychoakustische Verdeckungseffekte und die ausgeprägte Frequenzselektivität des Ohres damit erklärt.

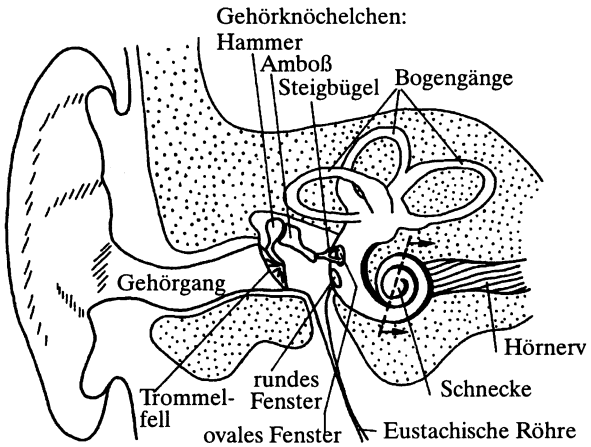


Abbildung 3.6 Schematische Darstellung des menschlichen Gehörs mit Außen-, Mittel- und Innenohr entnommen aus [202]. Das menschliche Gehörorgan besteht aus dem Außenohr mit Ohrmuschel und Gehörgang bis hin zum Trommelfell, dem Mittelohr vom Trommelfell bis hin zum ovalen Fenster und dem Innenohr mit der Schnecke und dem Cortischen Organ, das die Hörempfindung an den Hörnerv abgibt.

3.2.1 Übertragungsfunktion des Gehörs

Zur klassischen Modellierung der akustischen Wahrnehmung haben besonders die Arbeiten von Zwicker [202], [203], [204], Terhardt [172] und v. Békésy [178], [180] beigetragen. Zwischen dem Gehöreingang und dem eigentlichen Hörorgan im Innenohr befindet sich eine Übertragungskette, welche sich durch entsprechende Übertragungsfunktionen beschreiben läßt.

Der *äußere Gehörgang* kann näherungsweise als ein Rohr mit $l = 20\ldots25$ mm Länge und einem Durchmesser von $d = 8$ mm betrachtet werden, durch welches eine ebene Schallwelle läuft (wegen $\lambda_{\text{Schall}} \gg l$). Als Übertragungsfunktion des Gehörgangs $G(s)$ wird nach [172] das Verhältnis

$$G(s) = \frac{P_T(s)}{P_G(s)} \tag{3.28}$$

gewählt. $P_T(s)$ und $P_G(s)$ sind die Frequenzfunktionen des Schalldrucks am Trommelfell und am Eingang des Gehörgangs nach Laplacetransformation. Betrachtet man den Gehörgang als akustisches Zweitor, so läßt sich die Ketten-Übertragungsmatrix durch

$$\mathbf{A} = \begin{pmatrix} \cosh\left[\left(\delta + \frac{s}{c}\right)l\right] & \frac{\rho_0 c}{D}\sinh\left[\left(\delta + \frac{s}{c}\right)l\right] \\ \frac{D}{\rho_0 c}\sinh\left[\left(\delta + \frac{s}{c}\right)l\right] & \cosh\left[\left(\delta + \frac{s}{c}\right)l\right] \end{pmatrix}, \tag{3.29}$$

mit der frequenzunabhängigen Dämpfungskonstante δ, darstellen. Dabei sind c die Schallgeschwindigkeit und ρ_0 die Dichte der Luft bei Normalluftdruck und D die Querschnittsfläche des als Rohr modellierten äußeren Gehörgangs. $G(s)$ wird im hohem Maße durch die Abschlußimpedanz des Gehörgangs, also die Impedanz des Trommelfells, beeinflußt. Bezeichnet man die Trommelfeldimpedanz mit $Z_T(s)$, so ergibt sich mit der Kettenmatrix (3.29) die Übertragungsfunktion des Gehörgangs zu

$$G(s) = \frac{Z_T(s)}{Z_T(s)\cosh\left[\left(\delta + \frac{s}{c}\right)l\right] + \frac{\rho_0 c}{D}\sinh\left[\left(\delta + \frac{s}{c}\right)l\right]}. \tag{3.30}$$

Denkt man sich die Trommelfellseite mit einer unendlich großen Impedanz abgeschlossen, so ergibt sich mit (3.29) und (3.30) die Übertragungsfunktion des äußeren Gehörganges zu

$$G(s) \approx \frac{1}{\cosh\left[\left(\delta + \frac{s}{c}\right)l\right]}. \tag{3.31}$$

Der Betrag des Frequenzgangs nimmt bei ca. 3430 Hz ein Maximum an. Dies entspricht der als Gehörgangsresonanz bekannten Erscheinung, die sich durch ein Minimum der Hörschwelle bei dieser Frequenz bemerkbar macht, siehe Abbildung 3.9. In [172] wird für die *Übertragungsfunktion des Mittelohres* $M(s)$ folgende Beziehung angegeben

$$M(s) = \frac{C}{(s+\omega_g)(s+\omega_g)^2 + \omega_g^2}, \tag{3.32}$$

wobei C eine hörerabhängige Konstante und ω_g die Grenzfrequenz mit $\omega_g \approx 2\pi \cdot 1500\text{Hz}$ ist. Damit liegt Tiefpaßverhalten vor, was auch durch Untersuchungen in [178], [180], [202], [203] und [204] bestätigt wurde.

Die *Übertragungsfunktion des Innenohres* ist bisher noch unzureichend erforscht. Békésy konnte an Ohren von Leichen die Amplitudenverteilung längs des Cortischen Organs beobachten. Er fand heraus, daß die Transversalschwingungen der Basilarmembran eine Wanderwelle bilden. Dabei findet eine Frequenz-Orts-Transformation statt, siehe Abbildung 3.7.

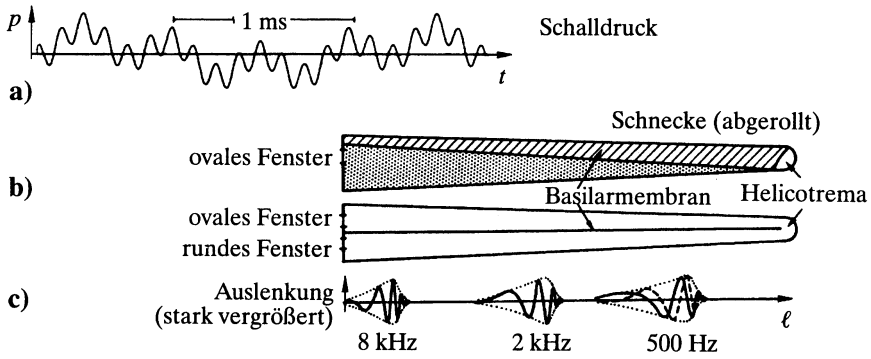


Abbildung 3.7 Schematische Darstellung der Frequenz-Orts-Transformation aus [202]. **(a)** Zeitsignal bestehend aus drei monofrequenten Komponenten gleicher Amplitude (500 Hz, 2000 Hz, 8000 Hz) **(b)** schematische Schnittdarstellung der abgerollten Cochlea **(c)** Wanderwellen-Resonanzen auf der Basilarmembran

3.2.2 Prothetische Aspekte des Hörens

Die Schallstärke, die das menschliche Gehör verarbeiten kann, liegt zwischen $10^{-5}...10^{2}$ Pa. Um in diesem Bereich nicht mit Exponenten rechnen zu müssen, werden Größen der Akustik häufig als logarithmische Größen in Pegeln [dB] gemessen. Als Bezugsgröße für den Schalldruck wird p_0 mit

$$p_0 = 20\mu Pa = 2 \cdot 10^{-5} \frac{N}{m^2} \tag{3.33}$$

festgelegt und für die Schallintensität gilt die Bezugsgröße I_0 mit

$$I_0 = 10^{-12} \frac{W}{m^2}. \tag{3.34}$$

Der Schallpegel L kann über die Schalldruck - oder Schallintensitätsverhältnisse durch

$$L = 10 \cdot \log \frac{I}{I_0} dB = 20 \log \frac{p}{p_0} dB \tag{3.35}$$

angeben werden.

3.2.2.1 Kritische Frequenzgruppe, Tonheit, Mithörschwelle

Das Gehör faßt in eng begrenzten Frequenzbändern Intensitäten von verschiedenen Schallreizen zusammen. Diese Frequenzbänder werden als *kritische Frequenzgruppen* bezeichnet. Reiht man über den gesamten Hörbereich alle kritischen Frequenzgruppen auf, so ergibt sich eine gehörorientierte nichtlineare Frequenzskala, die als *Tonheit* bezeichnet wird und die Einheit *Bark* besitzt (nach dem Dresdener Wissenschaftler *Barkhausen*). Sie stellt eine verzerrte Skalierung der Frequenzachse dar, so daß die Frequenzgruppen an jeder Stelle dieselbe Breite von genau 1 Bark haben. Daran angelehnt ist die aus der Sprachanalyse bekannte *Mel*-Skalierung, bei der gilt: 1 Bark = 100 Mel. Der nichtlineare Zusammenhang von Frequenz und Tonheit hat seinen Ursprung in der Frequenz-Orts-Transformation auf der Basilarmembran.

Die Tonheitsfunktion wurde von Zwicker [202] auf der Grundlage von Mithörschwellen- und Lautheitsuntersuchungen in Tabellenform angegeben. Es zeigt sich, daß im Hörfrequenzbereich von 0...16 kHz gerade 24 Frequenzgruppen aneinandergereiht werden können, so daß der dazugehörige Tonheitsbereich $z_k = 0 \ldots 24$ Bark beträgt. Eine empirische Formel, welche die tabellierten Werte gut annähert, lautet nach [172]:

$$\frac{f}{Hz} = 1960 \cdot \frac{\frac{z_k}{\text{Bark}} + 0.53}{26.28 - \frac{z_k}{\text{Bark}}} = \gamma(z_k). \tag{3.36}$$

Das Gehör ist in der Lage, an jeder Stelle f_c der Frequenzskala eine Frequenzgruppe mit der Breite von einem Bark zu bilden. Nach [172] kann für die Mittenfrequenz f_c die Bandbreite B_g der dazugehörigen Frequenzgruppe mit

$$B_g = 86 + 0.0055 \cdot f_c^{1.4} \tag{3.37}$$

approximiert werden. In [40] wurden mehrere Formeln angegeben, die den Zusammenhang zwischen der Frequenz f und der *Mel*-Skala f_M gut wiedergeben. Ein besonders einfaches Beispiel wird in (3.38) mit $a = 0.00024$ und $b = 0.471$ gezeigt:

$$f_M = \frac{f}{af + b}. \tag{3.38}$$

Die *Mithörschwelle* ist definiert als die Wahrnehmbarkeitsschwelle für einen Testschall bei gleichzeitiger Anwesenheit eines Störschalls. Liegt der Testschall unterhalb dieser psychoakustischen Schwelle, so liegt eine *Maskierung* vor.

Im Abschnitt 3.3 wird ausführlich auf diese Verdeckungseffekte eingegangen. Sie bilden die Grundlage für die in Abschnitt 6 entwickelten Verfahren zur Signalverbesserung und Geräuschreduktion und für die psychoakustische Systemidentifikation.

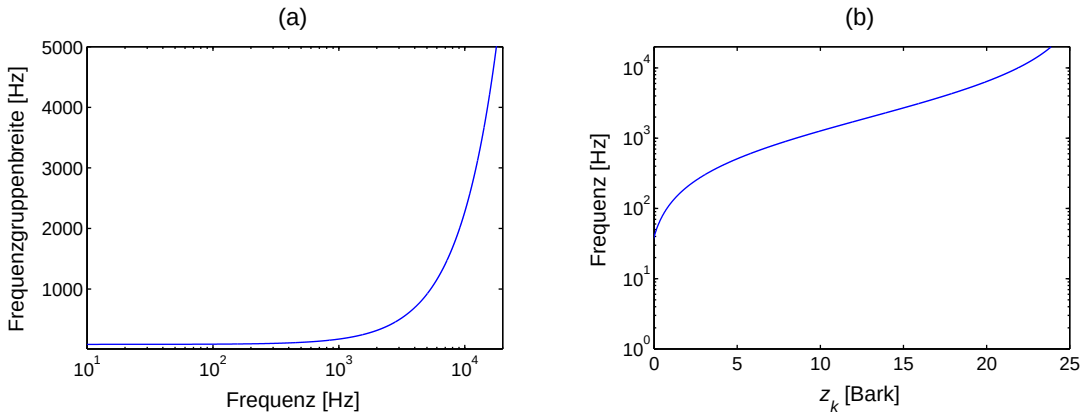


Abbildung 3.8 Kritische Frequenzgruppen. **(a)** kritische Frequenzgruppenbreite in Abhängigkeit von der Mittenfrequenz. Die Bandbreite beträgt in jedem Frequenzpunkt genau 1 Bark. **(b)** Abbildung der linearen Frequenzskala auf die psychoakustische Bark-Skala.

3.2.2.2 Wahrnehmung der Schallstärke, Lautheit

Die psychoakustische Empfindung der Lautstärke (*Lautheit*) unterscheidet sich von der wirklichen physikalischen Lautstärke. Die Lautheit in [*sone*] ist die subjektive Empfindung der Lautstärke bezogen auf einen 1-kHz-Ton mit einem Pegel von 40 dB. Die Steigerung des Lautstärkepegels um 10 phon entspricht der Zunahme der empfundenen Lautheit um den Faktor zwei.

Sinustöne verschiedener Frequenzen werden trotz gleichen Schallpegels unterschiedlich laut empfunden. Dieser Effekt wird in Abbildung 3.9 verdeutlicht. Der Lautstärkepegel (*phone* in [dB]) ist definiert als der Schalldruckpegel des als gleich laut empfundenen 1-kHz-Tones. Die Lautheitstheorie von Zwicker, vgl. [202], baut auf dem Schwellenmodell auf, dessen Basis die Verteilung der psychoakustischen Erregung entlang der Tonheit z_k ist, siehe Abschnitt 3.2.2.1.

Als mathematische Näherung der Kurven gleicher Lautheit $\Xi(f)$ wurde in [173] folgende Berechnungsvorschrift angegeben:

$$\Xi(f) = -0.6 \cdot 3.64 \cdot \left(\frac{f}{[kHz]}\right)^{-0.8} + 6.5 \cdot e^{-0.6\left(\frac{f}{[kHz]}-3.3\right)^2} - 10^{-3} \cdot \left(\frac{f}{[kHz]}\right)^4 . \tag{3.39}$$

Durch Abbilden der Formel (3.39) auf die Tonheitsskala z_k gemäß Gleichung (3.36) erhält man die Kurven gleicher Lautheit $\Xi(z_k)$, siehe Abbildung 3.9.

Zur Bildung der Erregungsverteilung aus dem Schallsignalspektrum wird der Verlauf der Mithörschwellen von Sinustönen bei Verdeckung durch Schmalbandrauschen zugrundegelegt. Dabei wird zwischen Kernerregung (innerhalb einer Frequenzgruppe) und Flankenerregung (außerhalb einer Frequenzgruppe) unterschieden.

Beispielsweise ist die psychoakustische Kernerregung eines Sinustones oder eines Schmalbandrauschens mit einer Bandbreite, die kleiner ist als die Frequenzgruppenbreite, gleich der physikalischen Schallintensität. Aus dem physikalischen Intensitätsdichtespektrum des eintreffenden zeitvarianten Schalls wird so die Verteilung der psychoakustischen Erregung $N'(z_k, t)$ gebildet.

Die Verteilung der psychoakustischen Erregung $N'(z_k, t)$ wird *spezifische Lautheit* benannt. Die Gesamtlautheit $N(t)$ ergibt sich als Integral über die spezifische Lautheit im Hörbereich (0...24 Bark) entlang der Tonheit-Skala z_k:

$$N(t) = \int_0^{24} N'(z_k, t) \frac{dz_k}{\text{Bark}}. \tag{3.40}$$

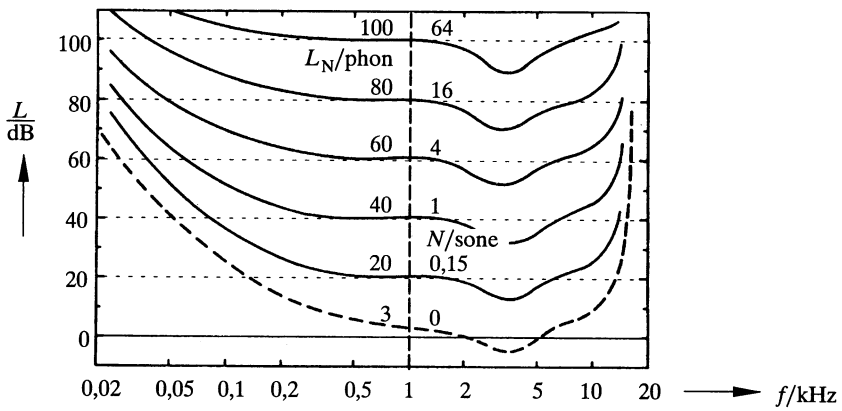


Abbildung 3.9 Kurven gleicher Lautheit von Sinustönen im ebenen Schallfeld. Schallpegel eines Sinustones, welcher die gleiche Lautheit hervorruft, wie ein 1 kHz-Ton mit angegebenen Pegel. Angaben in phon und sone. Grafik aus [202].

3.2.2.3 Differentielle Wahrnehmbarkeitsschwellen

Differentielle Wahrnehmbarkeitsschwellen sind im Unterschied zu Ruhe- oder Mithörschwellen Wahrnehmbarkeitsgrenzen für die Änderung einer Reizgröße. Hier sollen besonders die Wahrnehmbarkeitsschwellen für Phasen-, Amplituden- oder Frequenzänderung beachtet werden [202]. Als Faustformel gilt: *Amplitudenänderungen* sind dann wahrnehmbar, wenn sie innerhalb einer Frequenzgruppe 1 dB überschreiten. Für *Frequenzänderung*en liegt die Schwelle bei ca. 0.7% für Frequenzen oberhalb von 500 Hz und bei ca. 3,6 Hz für Frequenzen unterhalb von 500 Hz. Damit ist sie abhängig von der Frequenzgruppenbreite und beträgt ca. $1/27$ Bark. Diese empfindliche Schwelle ist über die Wanderwelle auf der Basilarmembran nicht zu erklären. Eine Änderung der Frequenz eines Sinustons um 1/27 Bark verschiebt die zugehöre Intensitätsverteilung derart, daß an der unteren Flanke in der benachbarten Frequenzgruppe eine Intensitätsänderung um 1 dB erfolgt. So wird eine Frequenzänderung als Amplitudenänderung in der benachbarten Frequenzgruppe wahrgenommen.

Generell ist bekannt, daß die Klangfarbe eines Sprachschalls im sehr geringen Maße von den *Phasenänderungen* der Teiltöne abhängt. Daher hat der genaue Umfang der Phasenänderung auf die Wahrnehmung des Sprachsignals keine Einfluß. Weil die Phase im Fall von komplexen Klängen aus vielen Teiltönen, wie sie zum Beispiel in Sprache oder Musik vorkommen, nur eine geringe Rolle spielt, kann man Sprach- und Musikklänge mit gewünschten Klangfarben erzeugen, indem man ohne Rücksicht auf die Phasen die entsprechenden Amplitudenspektren realisiert. Diese Eigenschaft wird häufig in Sprachsynthese- oder Geräuschreduktionssystemen ausgenutzt, da so die Phasenverläufe der Sprache explizit nicht nachgebildet werden müssen.

Die differentiellen Wahrnehmbarkeitsschwellen sind abhängig von vorhandenen Maskierern im Zeit- oder Frequenzbereich. Detaillierte Untersuchungen zu diesem Thema sind in [172] und [202] zu finden.

3.3 Psychoakustische Verdeckungseffekte

Psychoakustische Verdeckungseffekte sind die Grundlage für die in Abschnitt 6 entwickelten Verfahren der Sprachsignalverarbeitung und Geräuschreduktion. Aus der Praxis der Sprachkommunikation sind Maskierungseffekte bekannt. So werden z.B. ganze Wörter oder Wortsilben durch den Lärm auf einer belebten Straße oder in einer Werkhalle vollkommen bzw. teilweise verdeckt.

Auch innerhalb eines Audiosignals kann es zu Verdeckungseffekten kommen. Dabei sind die Informationen, die im psychoakustischen Verdeckungsbereich liegen, für das menschliche Gehör irrelevant. Etablierte Komprimierungs- und Datenreduktionsalgorithmen für Audiosignale nutzen diese Eigenschaft der Automaskierung von Audiosignalen aus, indem derartige

Informationsanteile reduziert werden können, ohne eine hörbare Qualitätsminderung des eigentlichen Signals zu bewirken.

Der umgekehrte Weg ist ebenfalls möglich: In jedes beliebige Audiosignal können zusätzliche nicht hörbare Informationen gebracht werden. Dabei wird durch psychoakustische Verdeckung dieser Informationen eine wahrnehmbare Störung des ursprünglichen Signals vermieden. So wurde in [132] ein Verfahren vorgestellt und patentrechtlich geschützt, bei dem das Lautsprecher-Raum-Mikrophon-System (LRM-System) durch orthogonale Testsequenzen identifiziert wird. Diese Anregungssequenzen werden so im Primärsignal versteckt, daß sie zwar im Empfänger ausgewertet werden können, doch für das menschliche Gehör nicht wahrnehmbar sind.

Zahlreiche Untersuchungen in [95], [123] und [202] haben gezeigt, daß Verdeckungseffekte für jedes menschliche Gehör bestimmbar sind. Individuelle Unterschiede treten kaum in Erscheinung. Prinzipiell kann man zwischen zwei Formen von Maskierern unterscheiden: *Simultane Verdeckung* im Frequenzbereich und *zeitliche Verdeckung* durch temporale Effekte entlang der Zeitachse. Außerdem existieren Mischformen dieser beiden Maskierungen.

In Abschnitt 6 werden drei neue Verfahren vorgestellt, die diese psychoakustischen Effekte nutzen. Dabei wurden klassische Geräuschreduktionssysteme so modifiziert, daß die Optimierung der Gewichtsfunktion nicht nur hinsichtlich der Minimierung des quadratischen Schätzfehlers erfolgt. Zusätzlich wurden psychoakustische Aspekte der menschlichen Wahrnehmung berücksichtigt, um eine adaptive Anpassung der Filterregel des Geräuschreduktionssystems an die Hörcharakteristik vorzunehmen. Dadurch konnte eine spürbare Verbesserung der resultierenden Sprachverständlichkeit und des subjektiven Höreindrucks erreicht werden.

3.3.1 Simultane Frequenzverdeckung

Bei simultaner Verdeckung treten Maskierer und Nutzsignal zur gleichen Zeit auf. Zur Untersuchung dieses Effektes wurden in [202] verschiedenen Hörern unterschiedlichen Alters und Geschlechts, Testsignale und Maskierer dargeboten. Ändert man nun die Form, Bandbreite, Amplitude und/oder Frequenz der Maskierer derart, daß die oft sinusförmigen Testsignale gerade hörbar werden, läßt sich die Mithörschwelle, siehe Abschnitt 3.2.2.1, bestimmen. Zwischen der Mithör- und der Ruhehörschwelle existiert der sogenannte *Verdeckungsbereich*, in dem das Testsignal nicht mehr hörbar ist.

3.3.1.1 Maskierung von Sinustönen durch weißes Rauschen

Wird einem Sinuston weißes Rauschen überlagert, so wird die Hörschwelle des Sinustones zur Mithörschwelle. Um die Wirkung des weißen Rauschens auf die Hörbarkeit eines Sinustones quantitativ zu beschreiben, wurde die Frequenz und Amplitude des Testsinustones über die gesamte Bandbreite des hörbaren Bereichs geändert und so die Mithörschwelle bestimmt,

siehe Abbildung 3.10. Eine wichtige Größe zum Kennzeichnen von frequenzabhängigen Rauschvorgängen ist die spektrale Schallintensität *I*, mit

$$I = \int_{f_u}^{f_o} \frac{\partial I}{\partial f} df . \tag{3.41}$$

Dabei wird ∂I als spektrale Schallintensitätsdichte bezeichnet. f_u und f_o stellen die untere und obere Grenzfrequenz des betrachteten Signals dar. Der spektrale Schalldichtepegel $l(f)$ kann mit

$$l(f) = 10 \cdot \log \frac{1}{I_0} \cdot \frac{\partial I}{\partial f} \text{ und } I_0 = 10^{-12} \frac{\mathrm{W}}{\mathrm{m}^2} \tag{3.42}$$

bestimmt werden.

Die Mithörschwelle eines Sinustones bei Verdeckung durch weißes Rauschen läßt sich folgendermaßen beschreiben: Unterhalb von 500 Hz liegt die Mithörschwelle des Sinustones ca. 17 dB oberhalb der Schallintensität des weißen Rauschens. Ab 500 Hz steigt dann die Mithörschwelle mit ca. 10 dB pro Dekade bzw. ca. 3 dB pro Oktave (Verdopplung der Frequenz) an.

Die Frequenzabhängigkeit der Mithörschwelle läßt sich mit der Frequenzgruppenbreite des Gehörs bei verschiedenen Mittenfrequenzen erklären:

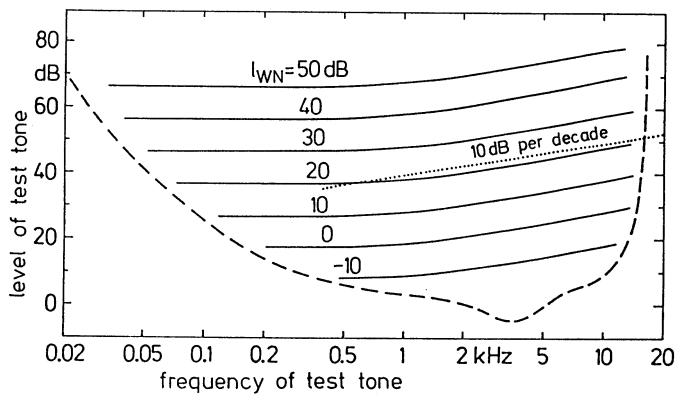


Abbildung 3.10 Maskierung eines sinusförmigen Testtones durch weißes Rauschen. Dargestellt ist die Schallintensität eines durch weißes Rauschen mit der Schallintensität l_{WN} gerade verdeckten Testtones in Abhängigkeit von seiner Frequenz. Grafik aus [202].

Weil das weiße Rauschen eine frequenzunabhängige Intensitätsdichte hat, erzeugt es in den einzelnen Frequenzgruppen des Gehörs einen um so stärkeren Verdeckungseffekt, je breiter die Frequenzgruppe ist. Deutlich ist in Abbildung 3.10 sichtbar, daß weißes Rauschen auch dann Verdeckung hervorruft, wenn es einen negativen Schallpegel hat. Der Dichtepegel l_{WN} ist mit Bezug auf die Bandbreite 1 Hz definiert worden. Da das Gehör aber mit Bandbreiten von 100 Hz bis 3000 Hz arbeitet, bewirken Dichtepegel mit $l_{\mathrm{WN}} < 0$ noch Verdeckungen.

3.3.1.2 Maskierung von Sinustönen durch schmalbandiges Rauschen

Bestimmt man die Mithörschwelle für schmalbandige Maskierer (Sinustöne, Schmalbandrauschen, frequenzgruppenbreites Rauschen), so zeigt sich, daß die spektrale Mithörschwelle gegenüber der Ruhehörschwelle auch dort angehoben ist, wo der Maskierer gar keine spektralen Anteile besitzt. Als Schmalbandrauschen wird frequenzgruppenbreites Rauschen benutzt, dessen Pegel mit L_{CB} bezeichnet wird.

In Abbildung 3.11 sind die Mithörschwellen von Sinustönen dargestellt, die durch frequenzgruppenbreites Rauschen mit der Mittenfrequenz 1 kHz und dem Schallpegel L_{CB} in Abhängigkeit von der Frequenz des Testtones verdeckt werden. Die Spitzen der Mithörschwellen steigen bei Erhöhung des Maskiererpegels um $\Delta L_{CB} = 20$ dB ebenfalls um 20 dB. Sie sind damit pegelunabhängig. Die unteren Flanken der Mithörschwellen, d.h. die in Richtung tiefer Frequenzen, besitzen eine vom Maskiererpegel praktisch unabhängige Steilheit von ca. -100 dB/Oktave. Diese große Steilheit wird an der oberen Flanke der Mithörschwellen nur für Maskiererpegel kleiner 40 dB erreicht. Bei größeren Pegeln wird die Flanke immer flacher und beträgt ca. -25 dB/Oktave bei $L_{CB} = 100$ dB.

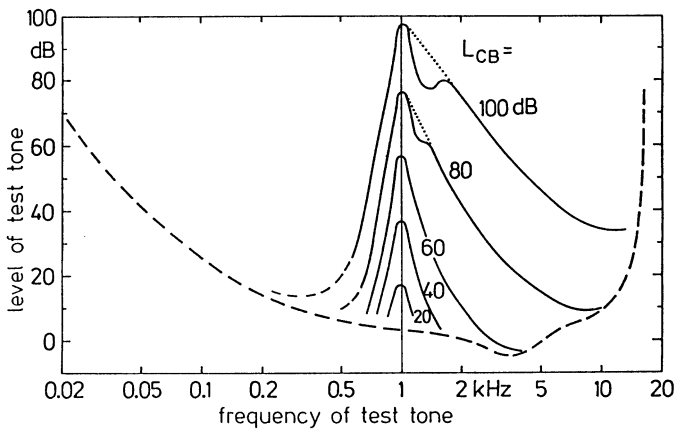


Abbildung 3.11 Maskierung im Frequenzbereich: Mithörschwellen bei einem Schmalbandrauschen mit der Mittenfrequenz $f_c = 1kHz$ und verschiedenen Pegeln L_{CB} als Maskierer und einem Sinuston mit der Frequenz f_T und dem Pegel L_T als Testschall. Die Ruhehörschwelle wurde gestrichelt eingezeichnet. Grafik wurde aus [202] entnommen.

Bei anderen Mittenfrequenzen als 1 kHz verhält sich das Gehör ähnlich. Die Flankensteilheiten von oberer und unterer Flanke sind nahezu unabhängig von der Mittenfrequenz des Maskierers, siehe Abbildung 3.12.

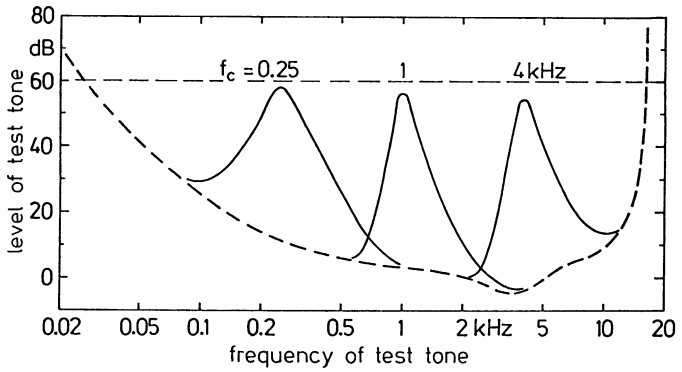


Abbildung 3.12 Mithörschwellen bei frequenzgruppenbreitem Schmalbandrauschen mit einem Pegel L_{CB} = 60 dB und Mittenfrequenzen von 250 Hz, 1 kHz und 4 kHz. Grafik aus [202] entnommen.

3.3.1.3 Maskierung von Sinustönen durch andere Sinustöne

Wird der sinusförmige Testton von einem weiteren Sinuston bei 1 kHz maskiert, ergeben sich in Abhängigkeit von der Frequenz des Testtones und des Pegels des Maskierers L_M Mithörschwellen, wie in Abbildung 3.13 dargestellt.

Wie schon im Abschnitt 3.3.1.2 beschrieben, ist die sogenannte Auffächerung der oberen Flanke in Abhängigkeit vom Pegel des Maskierers deutlich zu sehen, während die untere Flanke der Mithörschwelle nahezu frequenz- und pegelunabhängig ist. Für die obere Flankensteilheit lassen sich je nach Pegel des Maskierers -100...-25 dB/Oktave, für die untere Flankensteilheit ca. -100 dB/Oktave angeben. Zwischen Maskierpegel L_M und den Spitzen der Maskierschwellen L_T ergibt sich ein Offset von ca. 12 dB.

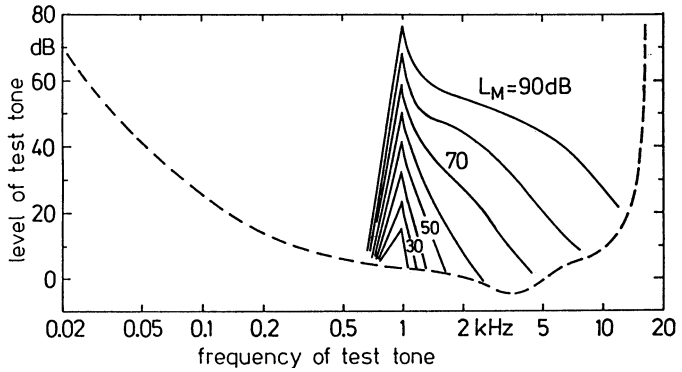


Abbildung 3.13 Verdeckung durch Sinustöne. Dargestellt ist die frequenzabhängige Mithörschwelle eines Testsinustones, der durch einen weiteren Sinuston mit der Frequenz 1 kHz und dem Pegel L_M maskiert wird. Grafik aus [202].

3.3.2 Zeitliche Verdeckung

Neben *Simultanverdeckung* tritt auch zeitliche Verdeckung auf. Dabei kann zwischen zwei Arten der zeitlichen Maskierung unterschieden werden: *Vorverdeckung* - schon vor Einschalten des Maskierers treten Verdeckungseffekte auf - und *Nachverdeckung* - nach Abschalten des Maskierers sinkt die Hörschwelle nicht sofort auf die Ruhehörschwelle. Vor- und Nachverdeckung sind schematisch in Abbildung 3.14 dargestellt und werden im Abschnitt 3.3.2.2 näher erläutert.

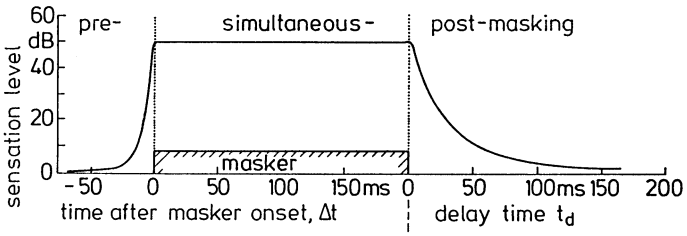


Abbildung 3.14 Schematische Darstellung von Simultan-, Vor- und Nachverdeckung. Dargestellt ist ein Maskierer und die jeweilige überhöhte Mithörschwelle. Deutlich ist sichtbar, daß die Mithörschwelle schon vor bzw. nach dem Maskierer über der Ruhehörschwelle liegt. Grafik aus [202].

3.3.2.1 Simultanverdeckung von Tonimpulsen

Die Ruhehörschwelle und die Mithörschwelle sind von der Dauer eines Testtones abhängig. Dabei können zwei unterschiedliche Effekte beobachtet werden: Die Abhängigkeit der Lautstärkeempfindung von der Dauer eines Testimpulses (siehe Abbildung 3.15) und der Zusammenhang zwischen Wiederholungsrate von kurzen Tonimpulsen und der Lautstärkeempfindung (siehe Abbildung 3.16). Nach [202] muß ein 20 ms-Impuls um 10 dB erhöht werden, um genauso laut wie ein 200 ms-Impuls empfunden zu werden.

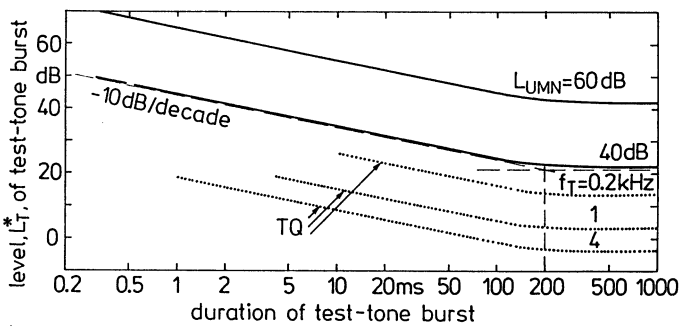


Abbildung 3.15 Abhängigkeit der Lautstärkewahrnehmung von der Dauer eines Testtonimpulses. Die gepunkteten Linien zeigen die Ruhehörschwelle in Abhängigkeit von der Dauer von Tonimpulsen unterschiedlicher Frequenz. Die durchgezogenen Linien stellen die Mithörschwelle in Abhängigkeit von der Dauer eines Sinustones dar. Dabei tritt weißes Rauschen als Maskierer mit dem Pegel L_{UMN} auf. Grafik entnommen aus [202].

Ab 200 ms Impulsdauer ist die Lautheit eines Tonimpulses unabhängig von seiner Dauer. Das Ohr integriert offensichtlich über die Dauer von Tönen kleiner als 200 ms. In Abbildung 3.16 ist die kleinere Flankensteilheit der Nachverdeckung im Vergleich zur Steilheit der Vorverdeckung deutlich sichtbar.

Nach Einschalten des Maskierers entsteht eine kurze Überhöhung der Mithörschwelle um etwa 10 dB. Dieser Effekt wird als *Overshoot*-Effekt bezeichnet und ist von der spektral unterschiedlichen Zusammensetzung von Testton und Maskierer innerhalb der kritischen Frequenzgruppen abhängig. Er verschwindet, sofern Maskierer und Testton gleiche spektrale Anteile besitzen.

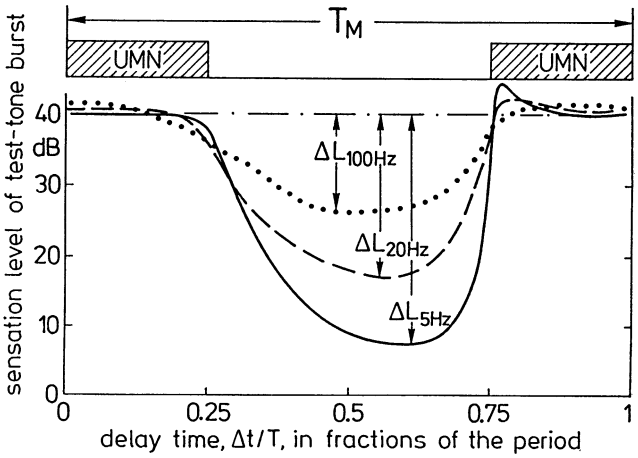


Abbildung 3.16 Abhängigkeit der Mithörschwelle von der Wiederholungsrate eines 3 kHz-, 3 ms-Testtonimpulses. Dabei liegt weißes, rechteckmoduliertes Rauschen mit den Modulationsfrequenzen 5, 20 und 100 Hz als Maskierer vor. Die Abzisse stellt die zeitliche Verschiebung des Testtonimpulses normiert auf die Periodendauer T_M des Maskierers dar. Die Strichpunktlinie zeigt die Mithörschwelle für einen ununterbrochenen 3 kHz Ton. Grafik aus [202] entnommen.

3.3.2.2 Vor- und Nachverdeckung

Ein Maskierer verdeckt den Testtonimpuls scheinbar, bevor er überhaupt eingeschaltet wird. Dies wird als *Vorverdeckung* bezeichnet. Eine Erklärung hierfür ist, daß laute Töne psychoakustisch schneller verarbeitet werden als leisere. Der Vorverdeckungseffekt ist wesentlich weniger ausgeprägt als der Nachverdeckungseffekt. Nach dem Abschalten des Maskierers sinkt die Hörschwelle nicht sofort auf die Ruhehörschwelle ab, sondern erreicht diese erst nach ca. 200 ms. Diesen Effekt bezeichnet man als *Nachverdeckung.* Er läßt sich mit dem langsamen Abschwingen der Wanderwelle auf der Basilarmembran erklären.

Auch die Bandbreite des Maskierers hat direkten Einfluß auf die Dauer der Nachverdeckung. Prinzipiell kann man davon ausgehen, daß in jeder separaten Frequenzgruppe Anteile des Maskierers Nachverdeckung entsprechend Abbildung 3.17 und Abbildung 3.18 hervorrufen.

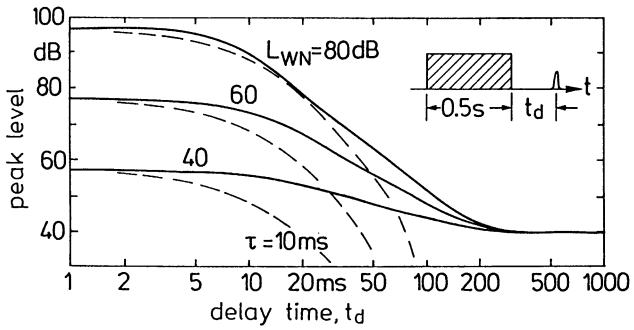


Abbildung 3.17 Nachverdeckung: Pegel eines gerade hörbaren 20 µs-Gaußimpulses als Funktion der Zeit nach Abschalten eines weißen Rauschmaskierers mit dem Pegel L_{WN}. Die gestrichelten Linien zeigen den exponentiellen Verlauf auf einer logarithmischen Pegel-Skala. Grafik entnommen aus [202].

Der Nachverdeckungseffekt ist auch abhängig von der Dauer des Maskierers. Dieser Umstand ist für die Sprachwahrnehmung von Bedeutung, da die meisten Laute kürzer als 200 ms sind.

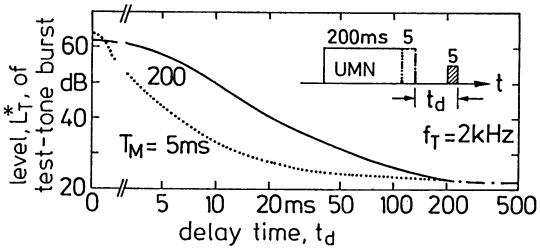


Abbildung 3.18 Nachverdeckung in Abhängigkeit von der Dauer des Maskierers. Die gepunktete Linie zeigt den Pegel eines gerade wahrnehmbaren 20 µs-Gaußimpulses bei Verdeckung mit Maskierer 5 ms Dauer und 60 dB Pegel. Die durchgezogene Linie beschreibt den Maskierer mit einer Dauer von 200 ms. Darstellung aus [202].

3.3.3 Additive Maskierung und Verdeckung durch komplexe Maskierer

Für die Ausnutzung der psychoakustischen Verdeckungseffekte in später vorgestellten Algorithmen und Verfahren stellt sich die Frage, ob diese Verhältnisse nur beim Vorliegen eines einzelnen Maskierers auftreten und welche resultierende Verdeckung sich bei zusammengesetzten komplexen oder mehreren additiv überlagerten Maskierern ergibt. Für die Simultanverdeckung können die Untersuchungsergebnisse aus [202] relativ klar zusammengefaßt werden. Für die zeitliche Verdeckung ist keine einfache Darstellung möglich. Allgemein zeigt sich, daß die Bandbreite der Maskierer eine wichtige Rolle spielt.

3.3.3.1 Addition simultaner Maskierer

Simultanverdeckung liegt vor, wenn mehrere Maskierer gleichzeitig auftreten. Abbildung 3.19 zeigt am Beispiel eines komplexen Tones, der aus einem 200 Hz-Ton und seinen neun Harmo-

nischen gebildet wurde. Der Grundton und die ersten vier Harmonischen liegen jeweils separat in verschiedenen Frequenzgruppen. Es kommt nicht zur additiven Überlagerung dieser Verdeckungsanteile. Die oberen Harmonischen liegen dagegen innerhalb einer Frequenzgruppe. Deutlich kommt es in dieser Frequenzgruppe zur additiven Überlagerung der einzelnen Mithörschwellen. Dabei entsteht ein ähnlicher Maskierungseffekt wie bei der durch schmalbandiges, frequenzgruppenbreites Rauschen hervorgerufenen Verdeckung. Dies entspricht auch den in Abschnitt 3.2.2.1 getroffenen Aussagen. Demnach kann die Addition von simultanen Maskierern nicht durch die Addition ihrer Intensitäten, sondern vielmehr durch Addition der einzelnen spezifischen Lautheiten modelliert werden. In Abschnitt 6.3.4 wird ein nichtlineares Modell eingeführt, das auf die Besonderheiten bei der Überlagerung von Maskierern eingeht.

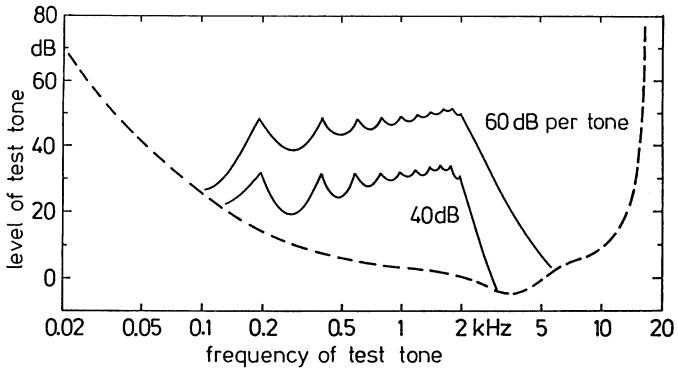


Abbildung 3.19 Simultanverdeckung durch einen komplexen Ton. Dargestellt ist Mithörschwelle für simultane Verdeckung eine sinusförmigen Testtones durch einen 200 Hz-Sinuston und neun Harmonische in Abhängigkeit von der Frequenz und des Pegels der Anregung [202].

3.3.3.2 Addition zeitlicher Maskierer

Die zeitliche Verdeckung ist von der Dauer und der Bandbreite des Maskierers abhängig. Im Vergleich zur Nachverdeckung spielt die Vorverdeckung eine sehr geringe Rolle für die Nutzung in späteren Algorithmen. Deshalb wird nachfolgend nur auf die Nachverdeckung eingegangen. Führt man eine zeitvariante spezifische Lautheit $N'_i(z_k, k)$ des i-ten Maskierers ein, so lassen sich die Verhältnisse bei Addition von simultanen Maskierern auch auf zeitliche Maskierer anwenden. Die spezifische Lautheit $N'_i(z_k, k)$ der Maskierer $i = 1, 2, \ldots$ wird dann zu jedem Zeitpunkt k innerhalb der Frequenzgruppen addiert. Aus der spezifischen Lautstärke wird die Mithörschwelle nach Abschnitt 3.2.2.1 gebildet, die dann innerhalb der Frequenzgruppen in ca. 200 ms auf die Ruhehörschwelle abschwingt.

3.3.3.3 Automaskierung

Als Automaskierung wird die Verdeckung eines Nutzsignals durch sich selbst bezeichnet. Der Automaskierung kommt im Abschnitt 6 besondere Bedeutung zu. Bei der auditiven Signalver-

besserung können so verfälschte Signalanteile unterdrückt oder modifiziert werden, die durch Schätzfehler der Störleistungsdichte bei der Anwendung der Geräuschreduktion entstanden sind. Deshalb liegt es nahe, nur in den hörbaren Bereichen oberhalb der Maskierschwelle eine Signalverarbeitung vorzunehmen. Unterhalb der Maskierschwelle ist keine weitere Geräuschreduktion erforderlich. Außerdem hat sich die Nutzung von Effekten der Automaskierung bei der auditiven Datenreduktion (z.B. MiniDisk, DCC) etabliert. Hierbei werden nur Informationen übertragen, die hörbar sind. Irrelevante Signale werden reduziert. Genauso lassen sich Signale unterhalb der Maskierschwelle verstecken. Diese Methode wird in [132] verwendet, um zusätzliche orthogonale Sequenzen in das Nutzsignal zu integrieren und so die verbesserten Korrelationseigenschaften des Signalgemisches für die schnelle und störsichere Systemidentifikation zu nutzen.

Automaskierung ist sowohl im Frequenzbereich wie auch im Zeitbereich zu beobachten. Abbildung 3.20 zeigt dafür ein Beispiel. Unterhalb der Mithörschwelle liegende Informationen können vernachlässigt werden, da sie vom menschlichen Gehör nicht wahrgenommen werden können. Setzt man diese Analyse in anderen Zeitrahmen fort, so läßt sich für das Gesamtsignal eine drastische Datenreduktion erreichen. Für die Automaskierung im Zeitbereich gelten analoge Aussagen.

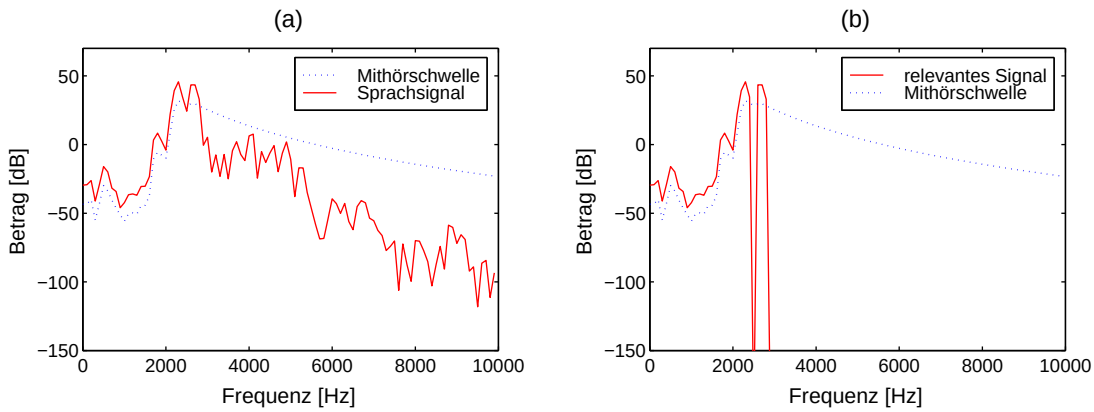


Abbildung 3.20 Automaskierung im Frequenzbereich. **(a)** Dargestellt sind Nutzsignal und Auto-Mithörschwelle für ein im Fahrzeug aufgenommenes Testsignal. Anschließend wurden die Mithörschwellen und der Verdeckungsbereich berechnet. Informationen unterhalb der Mithörschwelle werden durch das Nutzsignal selbst verdeckt und sind damit redundant **(b)**.

4 Überblick der Verfahren

Abbildung 4.1 zeigt die schematische Darstellung eines allgemeinen Geräuschreduktionssystems mit einem oder mehreren Mikrophonen[1]. Das grundsätzliche Ziel der folgenden Verfahren ist die möglichst fehlerfreie Schätzung $\ddot{s}(k)$ des gesprochenen zeitdiskreten Signals $s_0(k)$. Dafür steht nur das Signalgemisch

$$\mathbf{x}(k) = \left[x^{(1)}(k), x^{(2)}(k), \ldots, x^{(i)}(k), \ldots, x^{(I)}(k) \right]^T \tag{4.1}$$

in einem oder in I Mikrophonen zur Verfügung. Es setzt sich aus Sprache

$$\mathbf{s}(k) = \begin{bmatrix} s_0(k) * g_s^{(1)}(k) \\ s_0(k) * g_s^{(2)}(k) \\ \cdots \\ s_0(k) * g_s^{(i)}(k) \\ \cdots \\ s_0(k) * g_s^{(I)}(k) \end{bmatrix} = \left[s^{(1)}(k), s^{(2)}(k), \ldots, s^{(i)}(k), \ldots, s^{(I)}(k) \right]^T \tag{4.2}$$

und Störung

$$\mathbf{n}(k) = \begin{bmatrix} n_0(k) * g_n^{(1)}(k) \\ n_0(k) * g_n^{(2)}(k) \\ \cdots \\ n_0(k) * g_n^{(i)}(k) \\ \cdots \\ n_0(k) * g_n^{(I)}(k) \end{bmatrix} = \left[n^{(1)}(k), n^{(2)}(k), \ldots, n^{(i)}(k), \ldots, n^{(I)}(k) \right]^T \tag{4.3}$$

[1] Entsprechend der Anzahl der Mikrophone spricht man auch von ein-, bzw. mehrkanaligen Lösungen.

additiv zusammen, wobei zunächst Punktschallquellen für das Sprach- und Störsignal angenommen werden. Die vektorielle Darstellung verweist auf Mehrkanallösungen der Signalverarbeitung, bei denen sich in jedem Mikrophon unterschiedliche Anteile von Sprache und einfallender Störung ergeben. Hierbei sind $g_n^{(i)}(k)$ und $g_s^{(i)}(k)$ die Impulsantworten der Übertragungswege zwischen den Stör- und Nutzsignalquellen und den $i = 1, 2, \ldots, I$ Mikrophonen. Insgesamt ergibt sich eine verallgemeinerte vektorielle Darstellung für den am Eingang der adaptiven Signalverarbeitung zur Verfügung stehenden Signalvektor $\mathbf{x}(k)$ mit:

$$\mathbf{x}(k) = \mathbf{s}(k) + \mathbf{n}(k). \tag{4.4}$$

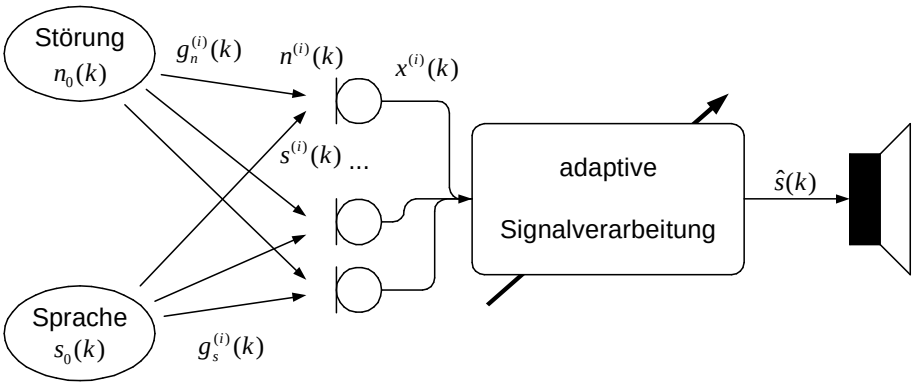


Abbildung 4.1 Allgemeine Darstellung von Geräuschreduktionssystemen. Zur Vereinfachung der Darstellung sind nur Direktschallanteile eingezeichnet. Schallreflektionen, Abschattungen und Signalverzögerungen gehen in die Raumimpulsantworten $g_s^{(i)}(k)$ und $g_n^{(i)}(k)$ ein.

Für die Systematisierung der Verfahren kann zum Beispiel die Zahl der verwendeten Mikrophone oder die Art der Modellbildung herangezogen werden. Aus den unterschiedlichen vereinfachten Modellannahmen ergeben sich Einschränkungen für die Anwendung der einzelnen Verfahren unter realen Bedingungen.

4.1 Einkanalige Geräuschreduktion

Da im allgemeinen keine a priori Kenntnisse von Signal und Störung vorliegen, steht bei Geräuschreduktionssystemen mit einem Mikrophon nur das unverzögerte gestörte Sprachsignal

$$x(k) = s(k) + n(k) \tag{4.5}$$

zur Verfügung. In zahlreichen Publikationen wurden Ansätze für die Lösung dieses Problems vorgestellt. Prinzipiell sind folgende Verfahren bekannt:

4.1.1 Spektrale Subtraktion

Als wichtigstes Verfahren hat sich die sogenannte *Spektrale Subtraktion* als eine Form der Wiener-Filterung etabliert, siehe Abschnitt 2.7. Dabei wird davon ausgegangen, daß Störung und Sprachsignal unkorreliert sind und sich additiv überlagern. Die spektrale Subtraktion hat sich vor allem als Vorverarbeitung für die automatische Spracherkennung in akustisch gestörter Umgebung [17] und auch für den Einsatz in Freisprecheinrichtungen [104] oder für das Entrauschen historischer Schallaufnahmen bewährt. Die Grundidee der spektralen Subtraktion liegt darin, den Betrag des Störspektrums zu schätzen und ihn vom Betrag des Spektrums des gestörten Sprachsignals zu subtrahieren. Die gestörte Phase wird dabei nicht verändert. Dabei wird von der Tatsache Gebrauch gemacht, daß das menschliche Gehör gegenüber Phasenverzerrungen weitgehend unempfindlich ist. Für die Spektrale Subtraktion ergibt sich folgende Problematik: Zur Schätzung der Störung steht nur das gestörte Sprachsignal im Mikrophon zur Verfügung. Mit Hilfe eines *Sprachpausendetektors* müssen deshalb die Zeitabschnitte bestimmt werden, in denen nur das Störsignal vorliegt. Durch Mittelung im Frequenzbereich über mehrere Pausensegmente wird ein Schätzwert für das Störspektrum in den nachfolgenden Sprachsegmenten gebildet. Dabei muß gefordert werden, daß das Störsignal stationären Charakter besitzt, damit der in der Sprachpause gewonnene Schätzwert noch während der nachfolgenden Sprachaktivität relevant ist. Dies wird in realen Anwendungen zu gravierenden Schwierigkeiten führen. Liegt beispielsweise ein Fehler in der Sprachpausendetektion vor, fließen Sprachsegmente in die Störschätzung mit ein, was sich in einer Verzerrung des Sprachsignals äußert. Bei nichtstationären Störgeräuschen ergibt sich ein zunehmender Fehler zwischen Schätzung des Störgeräusches und wirklicher Störung in den nachfolgenden Sprachsegmenten. Verbleibende Störgeräusche sind dann als sogenannte „*musical tones*“ wahrnehmbar. Sie sind eine Folge der im Frequenzbereich durchgeführten Operation der Subtraktion zweier Spektren. Hierbei ergeben sich einzelne, isolierte Spektrallinien, die sich nach der Rücktransformation in den Zeitbereich als kurzzeitige Sinustöne äußern.

In letzter Zeit sind Verfahren vorgeschlagen worden, die ohne explizite *Sprachpausendetektion* auskommen, z.B. [114], oder Reststörgeräusche durch *nichtlineare Filterung* reduzieren, vgl. [105]. Eine neue sehr effiziente Methode zur Signalverbesserung auf der Grundlage psychoakustischer Verdeckungseffekte wird in dieser Arbeit vorgestellt. Dabei wird ein Konzept zur Vermeidung von Restgeräuschen und Sprachverzerrungen erarbeitet und analysiert. Besonderes Augenmerk wurde dabei auf die Geräuschschätzung und auf die psychoakustische Modifikation der Wiener-Filter-Regel gerichtet.

4.1.2 Modellbasierte Verfahren

Modellbasierte Verfahren versuchen durch Nutzung von a priori zur Verfügung stehenden Kenntnissen bezüglich der statistischen Eigenschaften des Geräusch- und Sprachsignals oder durch Verwendung der Theorie der Spracherzeugung, Geräusch und Sprache zu trennen. So wurden in [33] und [164] die Nutz- und Störsignalparameter des Zustandraummodells für die *Kalman-Filterung* bestimmt. Dabei wird vom Modell der Spracherzeugung als autoregressiver Prozeß ausgegangen, vgl. Abschnitt 3.1.1.3. In [125] werden dagegen neuronale Netze genutzt, um die chaotischen Eigenschaften des Störgeräusches zu prädizieren.

Für die Verarbeitung gestörter Sprache in Spracherkennern hat sich die sogenannte *Kepstrale Lifterung* [139] bewährt. Hierbei wird die Dimension des gestörten Signalvektors im kepstralen Bereich reduziert. Ausgehend vom Fant'schen Source-Filter-Modell der Spracherzeugung, vgl. Abschnitt 3.1.1.1, lassen sich durch homomorphe Entfaltung Anregungs- und Vokaltraktanteile des Sprachsignals trennen. Nach der Lifterung liegen dann nur stimmhafte Anteile der Sprache vor, zusätzliche Geräusche werden entfernt. In diesem Zusammenhang haben auch die Techniken der *linearen Prädiktion* große Verbreitung gefunden. Dabei werden aus dem gestörten Sprachsignal die Modellparameter des AR-Prozesses der Spracherzeugung geschätzt. Durch anschließende Signalsynthese kann so das ungestörte Sprachsignal nachgebildet werden [198]. Diese Schätzung funktioniert aus den in Abschnitt 3.1.1.3 genannten Gründen besonders gut für stimmhafte Laute (z.B.: CELP-Codierung[1]).

Eine weitere Möglichkeit der modellbasierten Geräuschreduktion besteht darin, Modellparameter der akustischen Störung in sogenannten *Markov-Ketten* zu erfassen. Die so trainierten Markov-Modelle sind dann im realen Betrieb in der Lage, bei Eintreten bestimmter Störungen deren weiteren Verlauf vorherzusagen, vgl. [43], [50] und [102].

Allgemein läßt sich sagen, daß modellbasierte Verfahren immer dann zum Erfolg führen, wenn ausreichende Kenntnisse der Signale, Störungen oder des Systems vorliegen. In praktischen Anwendungen sind diese Bedingungen sehr schwer zu realisieren.

4.1.3 Nichtlineare Verfahren

Eine weitere Möglichkeit der Geräuschreduktion bieten nichtlineare Verfahren. Hierbei wird die beeindruckende Fähigkeit des Gehörs ausgenutzt, gegenüber nichtlinearen Verzerrungen relativ unempfindlich zu sein. Das bekannteste Verfahren in diesem Zusammenhang ist das sogenannte *Center-Clipper-Verfahren*, bei dem nur Signalanteile um den Mittelwert des gestörten Signals innerhalb eines Signalfensters weiterverarbeitet werden. Anteile über bzw. unter der Center-Clipper-Schwelle werden gesperrt (engl.: „clip“). Da Sprachsignal und Stö-

[1] CELP: Code Excited Linear Prediction

rung stets mittelwertfrei sind, kann so besonders in Sprechpausen eine deutliche Geräuschreduktion erreicht werden. Dies ist vergleichbar mit der Funktionsweise eines Kompanders, wobei eine adaptive Auslegung der Kompanderkennlinie in Abhängigkeit von der Sprachaktivität denkbar wäre, wie z.B. in [117] vorgestellt.

Eine weitere Möglichkeit der nichtlinearen Geräuschreduktion stellen die *Median-*, *Minimal-* und *Maximalfilterung* dar. Diese Verfahren kommen in der Meßtechnik zum Verwerfen von sogenannten „Meßausreißern“ zum Einsatz. Hierbei werden zum Zeitpunkt i je nach Analyselänge des Filters $N-1$ vergangene Meßwerte erfaßt und mit dem *i*-ten Meßwert sortiert. Je nach Art des Filters wird dann der *i*-te Meßwert durch das Maximum, Minimum oder den Median der Sortierfolge ersetzt. In [105] wurde dieses Verfahren in Verbindung mit der Spektralen Subtraktion eingesetzt, um verbleibende Reststörungen zu verringern.

Verfahren der nichtlinearen Glättung wurden erfolgreich bei der Signalnachverarbeitung angewandt, haben sich jedoch als autonome Methoden der Geräuschreduktion nur in wenigen Ausnahmefällen und speziellen Anwendungen (z.B. 2-D Bildverarbeitung) etabliert. Den größten Nachteil erfahren diese Verfahren durch die Verzerrung des Sprachsignals, was letztendlich zur einer Verringerung der Sprachverständlichkeit führt.

4.1.4 Multiratenverarbeitung, Filterbänke und Wavelets

Multiratenverarbeitung und Wavelets sind schon aus der Quellencodierung (Kompression) von Audio- und Videodaten bekannt. Dabei werden Oktavfilterbänke für die Analyse und Synthese des Signals genutzt. Eine besondere Bedeutung haben die *dyadischen Wavelets*. Das sind Transformationskerne, die für die Analyse nichtstationärer Signale mit Mehrfachauflösung, die sogenannte Wavelet-Analyse, verwendet werden. Es erfolgt so eine Transformation des nichtstationären Signals in den Zeit-Frequenzbereich, wobei zu verschiedenen Zeiten und bei unterschiedlichen Frequenzen eine variable Auflösung vorgenommen werden kann. Prinzipiell findet die Wavelet-Geräuschreduktion in drei Schritten statt: Zuerst wird eine Signalanalyse und Transformation in den *Zeit-Frequenzbereich* mit unterschiedlicher Auflösung durchgeführt. Dies wird in der Regel mit *dyadischen Oktavfilterbänken* erreicht. Anschließend werden die so gewonnenen Waveletkoeffizienten einer Schwellwertentscheidung unterzogen, d.h. einzelne Koeffizienten werden anhand einer globalen oder einer adaptiven Schwelle zu Null gesetzt. Danach erfolgt in der Synthesefilterbank eine Rekonstruktion aus den übriggebliebenen Waveletkoeffizienten. Im Idealfall wurden die Schwellen so gewählt, daß genau die Störanteile $n(k)$ des Signalgemisches $x(k)$ entfernt werden. Für die Schwellenwahl haben sich verschiedene Algorithmen etabliert, die entweder von a-priori-Kenntnissen der Störung ausgehen oder Modelle verwenden, die nur in speziellen Anwendungsfällen zutreffen. Darin zeigt sich auch bei diesen Verfahren die größte Schwäche. Liegen kaum oder keine Kenntnisse der

Eigenschaften von Nutzsignal und Störung vor, ist eine allgemeine Geräuschreduktion nicht erreichbar. Gute Ergebnisse mittels Wavelettechnik wurden aber bei der Rauschreduktion und „Entknackung“ historischer Schallaufnahmen erzielt, siehe [100] und [101].

4.2 Mehrkanalige Verfahren

Bei den mehrkanaligen Verfahren werden für die Signalverbesserung und Geräuschreduktion statistische Eigenschaften von Nutz- und Störsignal, raumakustische Kenntnisse des Übertragungskanals, der Anordnung der Mikrophone sowie der Position und der räumlichen Charakteristik der Signalquellen ausgenutzt. Dazu verwendet man zwei oder mehrere Mikrophone, die im Raum verteilt angeordnet sind. Generell weisen die mehrkanaligen Verfahren den Vorteil auf, durch räumliche Anordnung der Mikrophone und Kreuzkorrelation der Mikrophonsignale eine Aussage über den Charakter der Störung zu gewinnen. Damit können in einem Signalgemisch kohärente von inkohärenten Anteilen getrennt werden. Zudem kann die für die Leistungsdichtebestimmung notwendige Zeitmittelung mindestens teilweise durch die Scharmittelung der Mikrophonsignale ersetzt werden, was besonders für die Reduktion nichtstationärer Störsignale vorteilhaft ist.

4.2.1 Geräuschkompensation

Bei der Geräuschkompensation wird davon ausgegangen, daß in einem vom Nutzmikrophon räumlich getrennten Referenzmikrophon, nur das Störsignal vorliegt. 1975 wurde erstmals in geschlossener Form ein Störreduktionssystem beschrieben, das nach dem Prinzip der adaptiven Filterung funktioniert.

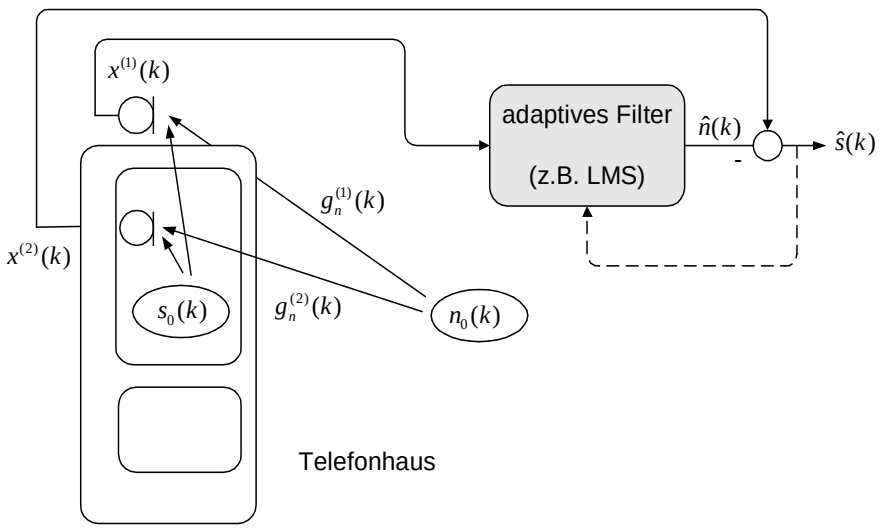


Abbildung 4.2 Geräuschkompensation mit zwei Mikrophonen. Zur Vereinfachung der Darstellung wurden die Raumimpulsanworten vom Sprecher zu den beiden Mikrophonen nicht eingezeichnet.

Das dem Nutzsignal überlagerte Störgeräusch wird durch Minimierung des mittleren quadratischen Fehlers (LMS-Algorithmus) geschätzt und danach durch Subtraktion verringert [193]. In Abbildung 4.2 ist ein Geräuschkompensationssystem, wie es z.B. in [150] realisiert wurde, dargestellt. Aus der allgemeinen Darstellung eines mehrkanaligen Geräuschreduktionssystems in Abbildung 4.1 und mit den Bezeichnungen der Eingangssignale zweier Mikrophone $x^{(1)}(k)$ und $x^{(2)}(k)$ sowie lt. Gleichung (4.4) folgt:

$$\mathbf{x}(k) = \begin{bmatrix} x^{(1)}(k) \\ x^{(2)}(k) \end{bmatrix} = \begin{bmatrix} s_0(k) * g_s^{(1)}(k) + n_0(k) * g_n^{(1)}(k) \\ s_0(k) * g_s^{(2)}(k) + n_0(k) * g_n^{(2)}(k) \end{bmatrix} = \begin{bmatrix} s^{(1)}(k) + n^{(1)}(k) \\ s^{(2)}(k) + n^{(2)}(k) \end{bmatrix}. \tag{4.6}$$

Abbildung 4.2 zeigt ein Beispiel für die Geräuschkompensation mit zwei Mikrophonen. Dabei wird ein Mikrophon als Referenzmikrophon mit dem Eingangssignal $x^{(1)}(k)$ und ein zweites Mikrophon als Sprachmikrophon mit dem Eingangssignal $x^{(2)}(k)$ eingesetzt. Das Kompensationsfilter wird so adaptiert, daß der mittlere quadratische Fehler

$$\begin{aligned} E\{\hat{s}^2(k)\} &= E\{[x^{(2)}(k) - \hat{n}(k)]^2\} \\ &= E\{[s^{(2)}(k) + n^{(2)}(k) - \hat{n}(k)]^2\} \\ &= E\{[s^{(2)}(k)]^2\} + E\{[n^{(2)}(k) - \hat{n}(k)]^2\} + 2E\{s^{(2)}(k) \cdot [n^{(2)}(k) - \hat{n}(k)]\} \\ &= E\{[s^{(2)}(k)]^2\} + E\{[n^{(2)}(k) - \hat{n}(k)]^2\} \end{aligned} \tag{4.7}$$

minimiert wird. Hierbei wurde vorausgesetzt, daß keine Sprachanteile $s^{(1)}(k)$ in das Referenzmikrophon gelangen und Sprache und Störung statistisch unabhängig sind. Sofort wird die Problematik deutlich: Nur unter der Annahme, daß die beiden Übertragungswege $g_n^{(1)}$ und $g_n^{(2)}$ nahezu identisch und beide Mikrophone möglichst entkoppelt sind (oder hinreichend weit auseinander liegen), kann eine genaue Kompensation der Störung vorgenommen werden. Das ist aber ein physikalischer Widerspruch. In der Realität zeigt sich, daß Sprachanteile in das Referenzmikrophon gelangen (Übersprechen) und eine Identifikation der verschiedenen Übertragungswege wegen der starken Instationarität der Störungen und des recht langsamen Adaptionsalgorithmus sehr schwierig ist.

Durch die Orthogonalisierung der Adaptionssignale mittels Lattice-Strukturen [150] kann die Kompensationsgeschwindigkeit erhöht werden. Der Einsatz von rekursiven adaptiven Filtern zur schnelleren Adaption und Geräuschkompensation hat sich aus Stabilitätsgründen hier nicht bewährt. In Gleichung (4.7) stellt das Sprachsignal für die Adaption des Kompensationsfilters eine nicht stationäre Störung dar. Die Kompensation kann deshalb nur sehr langsam erfolgen

oder muß in Sprachpausen stattfinden. Liegen mehrere oder expandierte Störquellen vor, verschlechtert sich der Kompensationsgewinn drastisch. Wegen der genannten Schwierigkeiten, haben sich derartige Verfahren für den Einsatz in Kraftfahrzeugen nicht durchgesetzt.

4.2.2 Adaptive und superdirektive Mikrophonarrays

Als adaptive Mikrophonarrays werden Mikrophonanordnungen aus mehreren Mikrophonen bezeichnet. Mit Hilfe adaptiver Signalverarbeitung (sog. *Beamforming*) wird der Ort des Sprechers in einem Raum geschätzt und die größte Empfindlichkeit des Mikrophonarrays (sog. „Keule“) automatisch auf seine Position gerichtet. Untersuchungen dazu wurden z.B. in [53], [62], [160] und [170] durchgeführt. Gerade für das sogenannte Beamforming-Verfahren ist die Erhöhung der Richtwirkung des Arrays nur mit weiteren Mikrophonen zu erreichen. Eine erweiterte Variante, z.B. in [121], nutzt die Nebenmaxima („Nebenkeulen“) der Empfindlichkeit des Mikrophonarrays aus, um ein Referenzsignal für die in Absatz 4.2.1 beschriebene Geräuschkompensation zu gewinnen.

Prinzipiell zeigt sich bei allen mehrkanaligen Realisierungen, daß die Fokussierung auf eine Nutzsignalquelle besonders in halligen Räumen oder bei starken zeitlichen Änderungen des akustischen Übertragungskanals sehr schwierig ist. Dadurch kommt es zu Adaptionsfehlern, die sich in hörbaren Signalverzerrungen und Pegelschwankungen äußern. Bei zeitinvarianten Übertragungswegen und diffusen Geräuschquellen wird dennoch im Vergleich zu einkanaligen Verfahren eine deutlich höhere Sprachverständlichkeit und Geräuschkompensation erreicht. Dabei muß aber der höhere Aufwand in der Systemarchitektur und Signalverarbeitung beachtet werden. In einigen Publikationen werden bis zu 16 Mikrophone in Anordnungen genutzt, die für praktische Anwendungen wie z.B. im Kraftfahrzeug zu aufwendig sind.

4.2.3 Kohärenzverfahren mit mehreren Mikrophonen

Liegt an zwei Mikrophonen ein Signalgemisch nach

$$\begin{aligned}\mathbf{x}(k) &= \begin{bmatrix} x^{(i)}(k) \\ x^{(j)}(k) \end{bmatrix} \\ &= \begin{bmatrix} s^{(i)}(k) + n^{(i)}(k) \\ s^{(j)}(k) + n^{(j)}(k) \end{bmatrix}.\end{aligned} \tag{4.8}$$

an, so ist die Kohärenzfunktion gemäß Gleichung (2.42) zwischen zwei stationären Mikrophonsignalen $x^{(i)}$, $x^{(j)}$ mit $i \neq j$ folgendermaßen definiert:

$$C_{xx}^{(ij)}(\Omega) = \frac{\left|R_{xx}^{(ij)}(\Omega)\right|^2}{R_{xx}^{(ii)}(\Omega) \cdot R_{xx}^{(jj)}(\Omega)}. \tag{4.9}$$

Dabei stellen $R_{xx}^{(ij)}(\Omega)$ das Kreuzleistungsdichtespektrum und $R_{xx}^{(ii)}(\Omega)$ sowie $R_{xx}^{(jj)}(\Omega)$ die Autoleistungsdichtespektren von $x^{(i)}(k)$ und $x^{(j)}(k)$ mit der normierten Frequenz Ω dar. Für ein homogenes diffuses Störschallfeld[1] hängt die Kohärenzfunktion der Mikrophonsignale nur noch von den Richtcharakteristika und dem Abstand der Mikrophone ab [113].

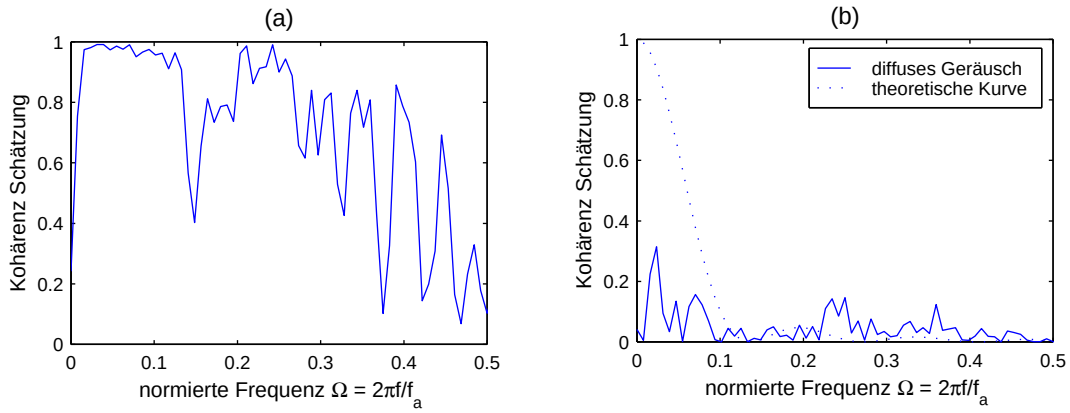


Abbildung 4.3 Schätzung der Kohärenz für zwei Mikrophone mit einem Abstand d = 10 cm und der Abtastfrequenz von f_a = 16 kHz. **(a)** reine Sprachprobe im PKW **(b)** Geräuschaufnahme während der Fahrt im BMW 528i touring. Beide Mikrophone wurden auf der Sonnenblende des Fahrers plaziert. Die gepunktete Linie zeigt die theoretische Kohärenz eines idealen diffusen Schallfeldes.

Geht man von einer diffusen Verteilung der Störer aus und stammt das Nutzsignal aus einer konzentrierten Quelle, so lassen sich die Korrelations- und Kohärenzverhältnisse in unterschiedlichen Mikrophonen nutzen, um Geräusch und Nutzsignal zu trennen, siehe Abbildung 4.3. Zahlreiche Autoren haben sich mit diesem Problem auseinandergesetzt. Beispiele sind in [25], [53], [89], [113] und [122] zu finden. Besonders erfolgreich konnte das Verfahren bei der Merkmalsextraktion von gestörter Sprache für den Einsatz in geräuschrobusten Spracherkennern eingesetzt werden [130].

Alle Verfahren treffen generell die Annahmen, daß eine konzentrierte Nutzsignalquelle in mehreren Mikrophonen kohärente Anteile erzeugt, wobei ein diffuses oder außerhalb des sogenannten Hallradius[2] liegendes Störgeräusch nur geringe Kohärenz in den verschiedenen Mikrophonen aufweist. Während die Reduktionsergebnisse in höheren Frequenzbereichen

[1] Das homogene diffuse Schallfeld zeichnet sich durch Schallwellen aus, die in allen Raumpunkten mit gleicher Intensität aus allen Raumrichtungen eintreffen.

[2] Der Hallradius ist abhängig von der Nachhallzeit der betrachteten Umgebung. Genaue Untersuchungen sind dazu in [113] zu finden.

recht gut ausfallen, kann im Frequenzbereich unterhalb von ca. 300 Hz das Stör- und Nutzsignal nur unbefriedigend separiert werden. Ursache ist die hohe Kohärenz beliebig diffuser Störquellen im unteren Frequenzbereich. Dieser Zusammenhang wird nachfolgend kurz erläutert: Es liege ein diffuses Schallfeld mit unendlich vielen im Raum verteilten, voneinander statistisch unabhängigen Punktschallquellen vor. Die Signale, die in einem derartigen diffusen Schallfeld an zwei unterschiedlichen Raumpunkten aufgenommen werden, zeichnen sich durch stochastische Phasenbeziehungen aus, wobei die Phase gleichmäßig in $\langle 0, 2\pi \rangle$ verteilt sei. Für die Kohärenzfunktion $C^{(ij)}(f)$ läßt sich unter Verwendung von Mikrophonen mit Kugelcharakteristik folgende Beziehung ableiten [113]. Mit der Schallgeschwindigkeit c gilt:

$$C^{(ij)}(f) = \frac{\sin^2\left(2\pi f \frac{d_{ij}}{c}\right)}{\left(2\pi f \frac{d_{ij}}{c}\right)^2} = \mathrm{si}^2\left(2\pi f \frac{d_{ij}}{c}\right). \tag{4.10}$$

Abbildung 4.3 zeigt die so theoretisch berechnete Kohärenz als gepunktete Linie. Charakteristisch ist, daß mit zunehmender Frequenz f und zunehmenden räumlichen Abstand der Mikrophone d_{ij} die Kohärenzfunktion $C^{(ij)}(f)$ rasch abnimmt und nur bei tiefen Frequenzen relativ hoch ist. Die gleichmäßige Ausbreitung tieffrequenter Signalanteile im Raum ist aus der Unterhaltungselektronik als *Subwoofer-Effekt* bekannt. Dabei können die Tieftonlautsprecher von den Hochtönern separiert irgendwo im Raum untergebracht werden und dennoch entsteht im Raum ein ausgewogenes Klangbild. Umgekehrt wird klar, daß wegen der hohen Kohärenz der Störquellen im unteren Frequenzbereich eine Separierung des Nutzsignals sehr schwierig ist. Die Folge sind übermäßige Verzerrungen des Nutzsignals und eine Verschlechterung der Verständlichkeit der Sprache. Gerade im unteren Frequenzbereich tritt aber in Fahrzeugen die größte Störleistung auf. Untersuchungen in [113] haben gezeigt, daß Kohärenzverfahren für den Einsatz in Kraftfahrzeugen deshalb nur bedingt einsetzbar sind.

Dennoch sind im allgemeinen die Mehrkanalverfahren den Einkanalverfahren überlegen, insbesondere beim Vorliegen instationärer Störsignale. Für den Einsatz im Kraftfahrzeug sind dagegen diese Verfahren nur bedingt geeignet. Oft rechtfertigt der erreichbare Qualitätsvorteil nicht den höheren Aufwand bei der Auslegung der Mikrophonarrays. Hohe zusätzliche Kosten und Einschränkungen im Bauraum des Kraftfahrzeugs sprechen gegen die Nutzung von mehr als zwei Mikrophonen. Deshalb werden mehrkanalige Ansätze und Verfahren nicht weiter betrachtet. Untersuchungen und Ergebnisse zu einkanaligen Verfahren lassen sich aber leicht auf mehrkanalige Verfahren ausweiten.

5 Bewertungsmethoden und Vergleich

Ziel der nachfolgend vorgestellten Verfahren und Algorithmen ist die Verbesserung des gestörten Nutzsignals. Gütekriterium ist dabei der Qualitätseindruck beim Nutzer. Seine subjektive Einschätzung der Qualität wird von vielen Einflüssen geprägt. Ein Vergleich zu anderen Systemen und die Validierung und Optimierung des Systems sind so nur sehr schwer möglich. Die *auditive* Beurteilung der Güte untersuchter Verfahren kann deshalb nur anhand einer breiten Testpopulation und unter verschiedenen Bedingungen durchgeführt werden. Diese Art der Qualitätsbestimmung ist sehr aufwendig und konnte in der vorliegenden Arbeit nur teilweise realisiert werden. Seit einigen Jahren werden große Bemühungen angestellt, um Möglichkeiten der technischen Signalauswertung zu schaffen und dabei ohne Beteiligung von Testpersonen *instrumentelle objektive* Maße zu bestimmen, die weitgehend dieselbe Qualitätseinstufung liefern wie ein aufwendig durchgeführter Hörtest. Die etablierten Verfahren der auditiven und instrumentellen Qualitätsbewertung wurden weitgehend von [16], [44], [183] und [186] zusammengefaßt. Die in dieser Arbeit verwendeten Qualitätsbewertungsverfahren werden in Abschnitt 5.2 und 5.3 kurz vorgestellt. Weitergehende Methoden, vor allem für die instrumentelle Gütebeurteilung von Geräusch- und Echounterdrückungssystemen, werden in [64] beschrieben.

5.1 Experimentalumgebung und -konfiguration

Bei Untersuchungen von gestörter Sprache geht man meist von einer additiven Überlagerung von Nutz- und Störsignal aus. Diese Voraussetzung wird auch bei den theoretischen Herleitungen der Geräuschreduktions- und Schätzalgorithmen verwendet.

Die realistische Methode der Störüberlagerung ist jedoch die gemeinsame Aufnahme von Störung und Nutzsignal. Diese Methode hat zwei Nachteile:

(1) Die Aufnahme der Störsituation muß bei unterschiedlichen Signal-Störverhältnissen für jeden Meßpunkt einzeln erfolgen. Dadurch ergibt sich ein hoher Aufwand zur Durchführung der Aufnahmen. Außerdem ist die Reproduzierbarkeit einzelner Messungen nur eingeschränkt möglich.

(2) Die Untersuchung und die Bewertung der Algorithmen wird deshalb erschwert, da Nutzsignal und Störung nicht getrennt vorliegen und für den Vergleich aus der gestörten Sprachprobe geschätzt werden müssen. Die Berechnung des Signal-Störverhältnis ist besonders bei nichtstationären Störsignalen und Sprachproben äußerst schwierig.

Damit die Störsignaldämpfung und die Verbesserung des Sprechersignals im einzelnen gemessen werden können, müssen die betreffenden Signale, also das gestörte, das ungestörte Sprechersignal und die Störsignale, am Eingang und am Ausgang des Geräuschreduktionsverfahrens getrennt vorliegen und getrennt verarbeitet werden. Systematisch bedingte Einschränkungen, z.B. Nichtlinearitäten des hochwertigen Mikrophones, Quantisierungseffekte der A/D-D/A-Wandlung und Einflüsse durch den Interpolationstiefpaß sowie System- und Verstärkerrauschen werden nicht weiter untersucht, sondern werden als zusätzliche additive Fehler in das Geräuschsignal $n(k)$ modelliert. Das Sprachsignal $s(k)$ und das Störsignal $n(k)$ werden unter möglichst identischen Bedingungen separat aufgenommen. Das gestörte Sprachsignal $x(k)$ wird durch Summation des Sprecher- und Störsignals gebildet.

Das Eingangs-SNR wird eingestellt, indem ausschließlich die Verstärkung ξ der Störsignale variiert wird. Nutzsignal $s(k)$ und Störung $n(k)$ wurden folgendermaßen additiv überlagert

$$x(k) = \gamma \cdot \left[s(k) + \xi \cdot n(k) \right]. \tag{5.1}$$

Für die Einstellung des gewünschten SNR von $x(k)$ erhält man für ξ mit $k = 1, 2, \ldots, N$

$$\xi = \frac{\sum\limits_{m=aktiv} s^2(k,m)}{\sum\limits_{m=aktiv} n^2(k,m)} \cdot 10^{\frac{SNR_x}{10}}. \tag{5.2}$$

In Gleichung (5.2) werden nur Signalsegmente mit Sprachaktivität ($m = aktiv$) zur Leistungsberechnung verwendet. Diese Vorgehensweise ist an die Berechnung des segmentalen SNR, siehe Abschnitt 5.3.1, angelehnt. So ergibt sich ein Signal-Störverhältnis des Signalgemisches, das der wirklichen psychoakustischen Wahrnehmung im Gehör entgegenkommt. Durch den Faktor γ wird das Signalgemisch $x(k)$ auf die Amplitude $[-1, 1]$ normiert. Man erhält für γ mit $k = 1, 2, \ldots, N$:

$$\gamma = \frac{1}{\max\left(\left| s(k) + \xi n(k) \right| \right)}. \tag{5.3}$$

In Abbildung 5.1 ist die verwendete Experimentalumgebung dargestellt. Dabei wird stets von realen Signalproben, die im Pkw oder im Tonstudio aufgenommen oder synthetisiert wurden, ausgegangen. Für die Aufnahmen wurde ein Mikrophon vom Typ AGK-Q400 mit Hyperniercharakteristik verwendet. Das Mikrophon wurde mittig auf die Sonnenblende des Fahrers plaziert, wobei sich ein Abstand zum Fahrer von ca. 35 cm ergab. Als Aufzeichnungsgerät wurde ein DAT-Recorder genutzt.

Für die Auswertung und den Vergleich verschiedener Verfahren und Parameter wurde bei allen Simulationen und Messungen eine Abtastrate von $f_a = 11025$ Hz und eine Quantisierung von $q = 16$ Bit vorgegeben. Für die Analyse und Simulation wurden die Softwarepakete Matlab® und DaDisP® eingesetzt. Für alle Experimente wurden dieselben separierten Nutz- und Störsignale verwendet.

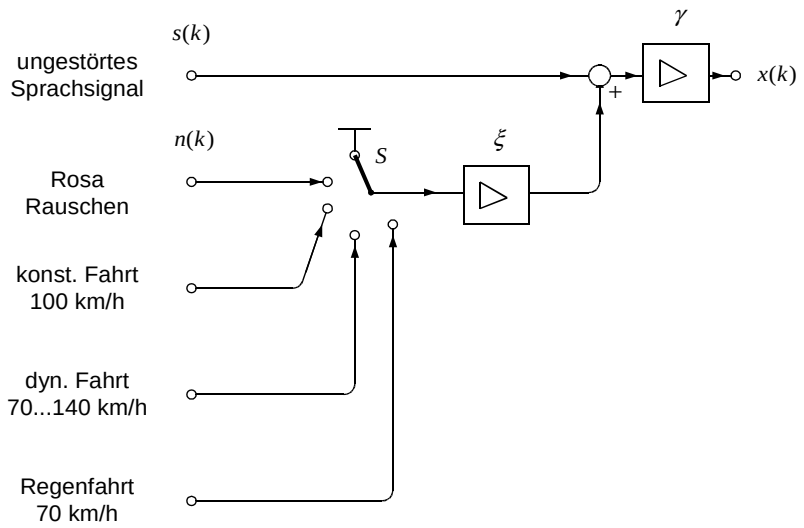


Abbildung 5.1 Experimentalumgebung und Simulationskonfiguration. Der Schalter S ermöglicht das Mischen unterschiedlicher Störsignale $n(k)$ mit dem Nutzsignal $s(k)$. Dabei wirkt der Verstärkungsfaktor ξ nur auf die Störung. Mit ihm wird das gewünschte SNR eingestellt. Der Faktor γ ermöglicht die Amplitudenormierung von $x(k)$.

5.1.1 Sprachsignale

Zunächst sind in Abbildung 5.2 und Abbildung 5.3 die Zeitverläufe und Spektrogramme der ungestörten Sprachproben einer männlichen und weiblichen Stimme dargestellt. Für die Sprachprobe wurde die Äußerung „eins-zwei-fünfundzwanzig“ gewählt, wobei zwischen „eins“ und „zwei“ bewußt eine Pause gelassen und dagegen bei „fünfundzwanzig“ fließend gesprochen wurde. Deutlich ist die Formantstruktur sowohl in der weiblichen wie auch in der männlichen Stimme zu erkennen. Die Signale wurden mit 11025 Hz abgetastet und mit 16 Bit Auflösung quantisiert.

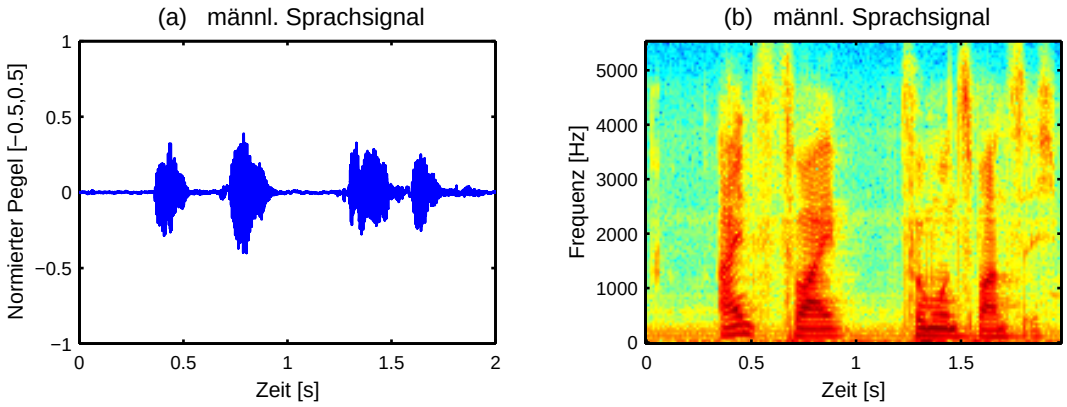


Abbildung 5.2 Reines Sprachsignal. Männliche Stimme. **(a)** Zeitverlauf **(b)** Spektrogramm mit FFT-Länge 256 und Overlap 128. Aufnahme im BMW 528i touring im Stillstand mit SNR = 60 dB .

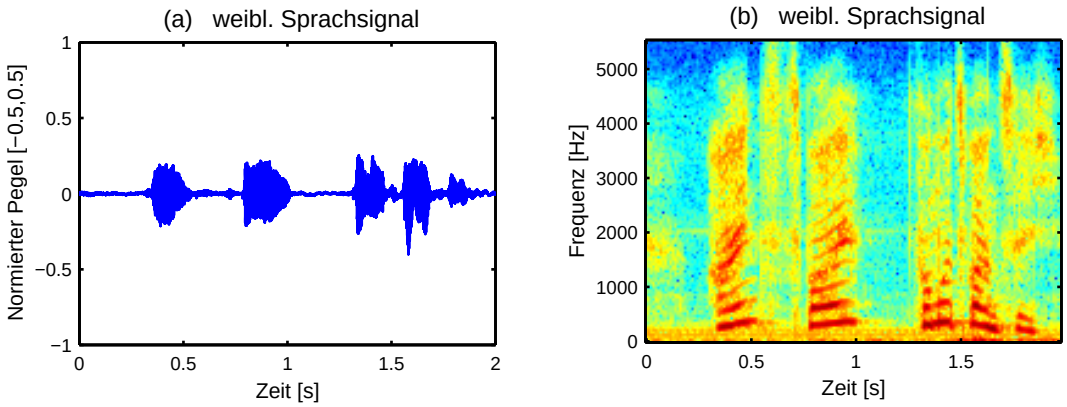


Abbildung 5.3 Reines Sprachsignal. Weibliche Stimme: „Eins-Zwei-Fünfundzwanzig“ **(a)** Zeitverlauf **(b)** Spektrogramm mit FFT-Länge 256 und Overlap 128. Aufnahme im BMW 528i touring im Stillstand mit SNR = 60 dB . Erklärung der Farben: rot: hohe Leistungsdichte, blau: geringe Leistungsdichte.

5.1.2 Störsignale

In den folgenden Abbildungen werden Geräusche und Störsignale dargestellt, wie sie in einer realen Testumgebung auftreten. Entsprechend Abbildung 5.1 werden Sprach- und Störsignale additiv überlagert, so daß die gestörten Sprachsignale für verschiedene Testfälle und gleichzeitig die separaten Störungen für den Systemvergleich zur Verfügung stehen.

In Abbildung 5.4 ist ein Beispiel eines Fahrgeräusches dargestellt, wie es bei konstanter Autobahnfahrt mit 100 km/h im Fahrzeug aufgenommen wurde. Besonders deutlich ist die hohe Leistungsdichte im unteren Frequenzbereich sichtbar. Das Störspektrum ändert sich nur sehr langsam mit der Zeit. Dagegen ist in Abbildung 5.5 das Fahrgeräusch für dynamische Fahrt und Beschleunigung auf der Landstraße von ca. 50 km/h auf 70 km/h zu sehen.

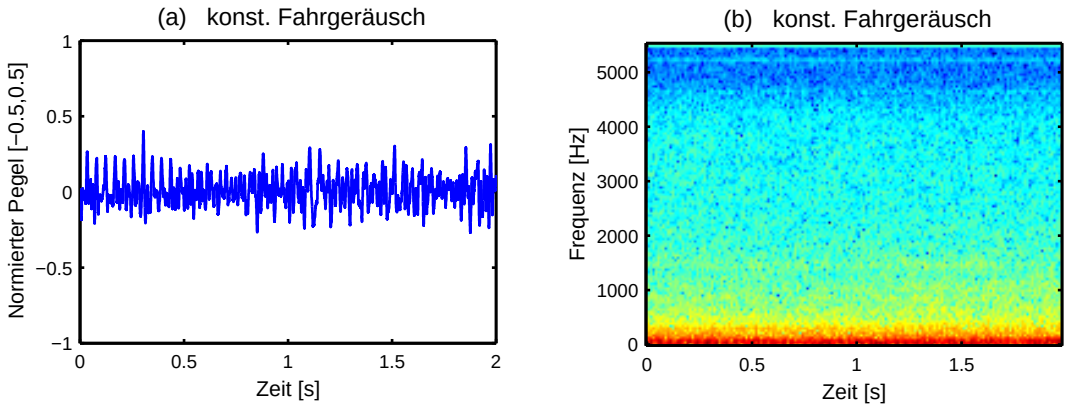


Abbildung 5.4 Fahrgeräusch im BMW 528i touring bei konstanter Autobahnfahrt mit 100 km/h.

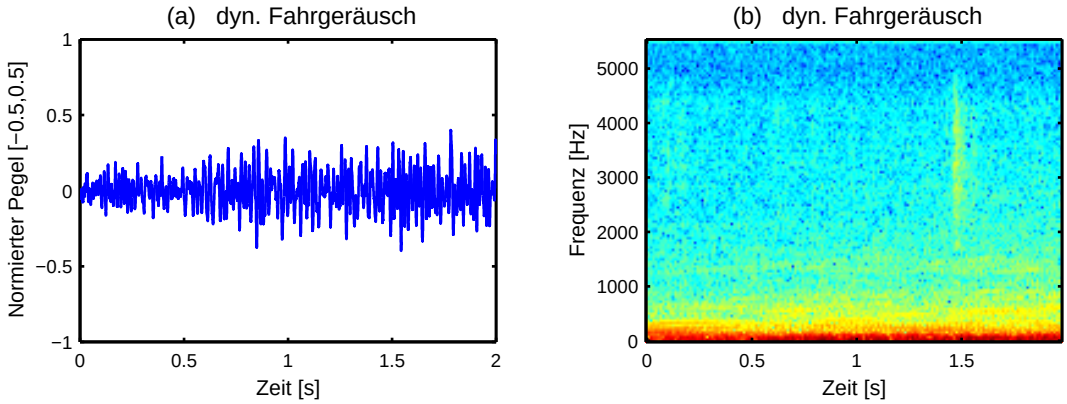


Abbildung 5.5 Dynamische Fahrt auf der Landstraße im BMW 528i touring. Dargestellt ist der Zeit- und Frequenzverlauf des Fahrgeräusches während der Beschleunigung von ca. 50 km/h auf 70 km/h.

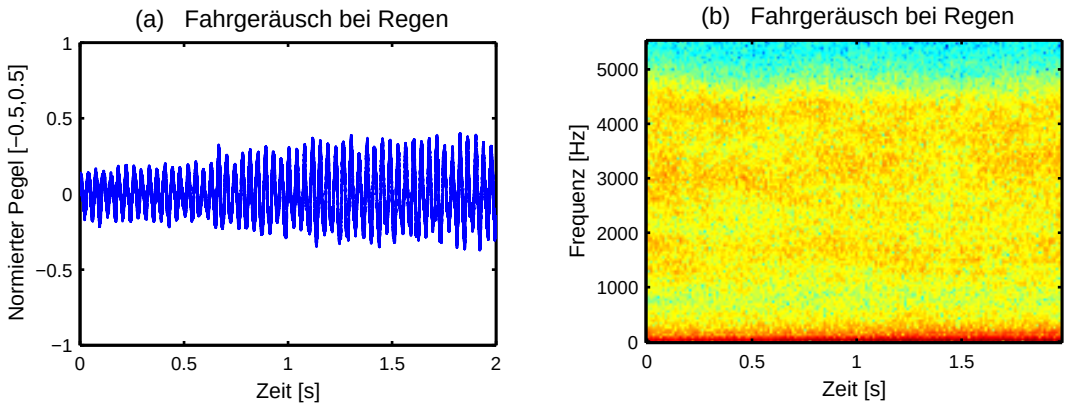


Abbildung 5.6 Geräusche bei Fahrt auf regennasser Landstraße mit ca. 70 km/h. Im mittleren Frequenzbereich ist eine starke zeitliche Varianz der Leistungsdichte des Fahrgeräusches zu erkennen.

Auch hier wurde eine hohe Leistungsdichte im unteren Frequenzbereich bis ca. 300 Hz festgestellt. Markant ist der zunehmende höherfrequente Anteil im Verlaufe der Beschleunigung nach ca. 1,5 Sekunden. Durch das Schalten in den höheren Gang ändert sich der Frequenzverlauf dann abrupt.

In Abbildung 5.6 ist der Signal- und Frequenzverlauf bei Fahrt auf regennasser Landstraße mit ca. 70 km/h dargestellt. Erneut zeigt sich bei tiefen Frequenzen die relativ hohe Leistungsdichte des Fahrgeräusches. Dennoch verursacht das aufgewirbelte Spritzwasser eine deutliche Verteilung der Geräuschleistung auch zu hohen Frequenzen hin. Das Fahrgeräusch ist im mittleren Frequenzbereich stark instationär.

In Abbildung 5.7 wird ein Beispiel eines rosa Rauschprozesses gezeigt. Charakteristisch für rosa Rauschen ist die besondere spektrale Verteilung der Intensität des Rauschsignals. Beim rosa Rauschen verringert sich die Intensität des Rauschen jeweils um 3dB pro Oktave. Dadurch ergibt sich ein abfallendes Leistungsdichtespektrum, wie in Abbildung 5.7 dargestellt. Rosa Rauschen wird als gleichmäßige rauschartige Erregung wahrgenommen.

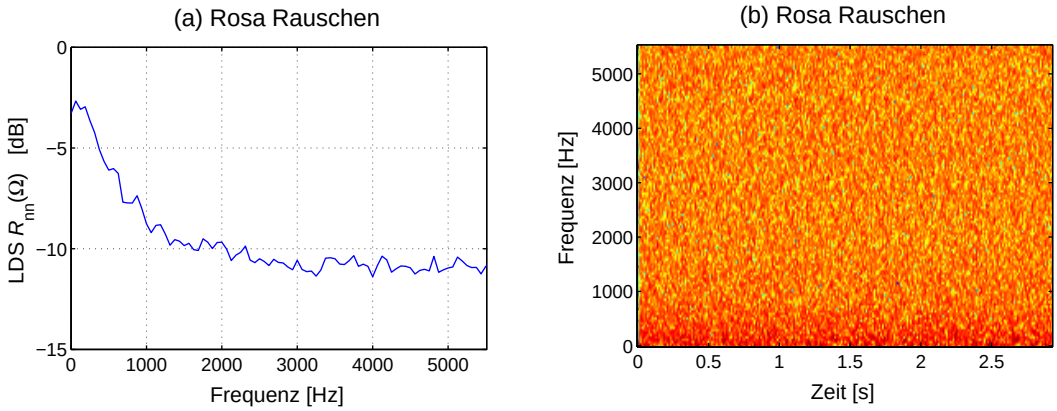


Abbildung 5.7 Mit Matlab® synthetisiertes rosa Rauschsignal. **(a)** Langzeit-LDS **(b)** Spektrogramm mit FFT-Länge 256 und Overlap 128.

5.2 Auditive subjektive Bewertung

Ein verarbeitetes Signal kann man durch Hörtests beurteilen, indem man es:

- mit einer Wertung belegt, die über mehrere Personen gemittelt wird (*Mean Opinion Score*),
- mit dem gestörten Nutzsignal vergleicht und die Qualitätsminderung bewertet oder
- mit dem ungestörten Nutzsignal vergleicht und die Qualitätsminderung einschätzt,
- mit einem zweiten, möglicherweise mit einem durch weitere Verfahren verarbeiteten Signal vergleicht (*Vergleichstest*)
- oder durch vorgegebene Attribute beschreibt (*Diagnostic Acceptability Measure*).

5.2.1 Mean-Opinion-Score

Für die Untersuchungen in dieser Arbeit wurde ein Vergleichstest gewählt, bei dem das geräuschreduzierte Signal durch zehn männliche und zehn weibliche Hörer zwischen 22 und 40 Jahren bewertet wird. Dabei wird der Unterschied zu anderen Geräuschreduktionsverfahren mit den Maßstäben nach Tabelle 5.1 bewertet. Durch Mittelung der einzelnen Bewertungen ergibt sich der Mean-Opinion-Score (*MOS*). Die Tests wurden im stehenden Fahrzeug mit geschlossenen Fenstern durchgeführt. Das Signal-Stör-Verhältnis ist frei konfigurierbar. Dabei wurden unterschiedliche Sprach- und Geräuschproben zur Verfügung gestellt. Die einzelnen Bewertungskriterien und deren Gewichtung für die Gesamtbewertung sind Tabelle 5.1 zu entnehmen:

Note	**Verständlichkeit** **60%**	**Verzerrung des Sprachsignals** **20%**	**Charakter der Reststörungen** **20%**	**Reduktions-ergebnis** **gesamt ∅**
5	kaum verständlich	sehr stark störend	starke musical tones und Restgeräusche	nicht akzeptabel
4	sehr schwer verständlich, raten	störend und unnatürlich	musical tones und starkes Restrauschen	schlecht
3	noch gerade verständlich	hörbar	hörbar	annehmbar
2	gut verständlich	teilweise kleine Verzerrungen wahrnehmbar	kaum hörbare musical tones, Reststörungen	gut
1	alles deutlich verständlich	keine wahrnehmbaren Verzerrungen	keine wahrnehmbaren musical tones, angenehme Reststörung	ausgezeichnet

Tabelle 5.1 Maßstäbe für die *Mean-Opinion-Score* Bewertung in Abhängigkeit von der Verständlichkeit und den Verzerrungen des Sprachsignals, dem Charakter der Reststörungen und dem subjektiven Gesamteindruck beim Testhörer.

5.2.2 Anker-Beurteilung, MNRU-Test

Die gemeinsame Auswertung getrennt durchgeführter Tests ist problematisch: Je nach Sprache, Sprecher- und Hörerauswahl und Erwartungshaltung der Hörerschaft divergieren die absoluten Urteile über gleiche Verfahren und Systeme. Die nötige Normierung gelingt, indem man bekannte Systeme in alle Tests integriert, deren Qualitäten den gesamten MOS-Bereich

von $MOS = 1\ldots5$ abdecken. Üblicherweise kommt hierfür der *Modulated-Noise-Reference-Unit-Test* (*MNRU-Test*) [88] zum Einsatz. Hierbei wird additives bandbegrenztes (rosa) Störrauschen erzeugt, dessen Stärke proportional zur Signalamplitude ist. Die Variationen der beiden Verstärkungsfaktoren γ und ξ in Abbildung 5.1 erlauben die Einstellung eines gewünschten SNR. Für die Einstellung des SNR von 40 dB ergibt sich beim MNRU-Test ein bewertungswert von $MOS = 4.4$. Für ein SNR von 10 dB erhält man $MOS = 1.9$. Für die Skalierung der einzelnen subjektiven MOS-Bewertungen wird deshalb stets eine Signalprobe mit rosafarbigem Testsignal gewählt.

5.3 Instrumentelle objektive Bewertung

Instrumentelle Qualitätsmaße bieten einen direkten Vergleich mit anderen Verfahren. Sie stellen ein objektives Kriterium für die Abschätzung der Güte dar. Problematisch ist die Diskrepanz zwischen objektiver Beurteilung und subjektivem Eindruck. Beispielsweise sagt die Verbesserung des Signal-Stör-Verhältnisses (*SNR*) nichts über die erreichte Verständlichkeit und hörbare Qualität einer Geräuschreduktion aus. Oft wird eine „angenehme“ Reststörung weniger störend empfunden als hörbare Verzerrungen der Sprache oder der Störung.

Dennoch liefern die objektiven Verfahren erste Anhaltspunkte für die Güte eines Verfahrens. Wegen ihrer einfachen Beschreibung und Implementation werden sie häufig für die Systemverifikation und -optimierung eingesetzt.

5.3.1 Segmental Signal-to-Noise-Ratio-Improvement (SNRI)

Von anderen nachrichtentechnischen Anwendungen ist das sogenannte *globale SNR* bekannt. Das globale SNR hat jedoch bezüglich der beim Hören empfundenen Qualität nur wenig Aussagekraft. Nur im Bereich relativ hoher Störabstände bewirkt die Erhöhung des SNR die Verbesserung des subjektiven Höreindrucks. Es ist folgendermaßen definiert:

$$SNR = 10\log\left(\frac{\sum_{k=1}^{N} s^2(k)}{\sum_{k=1}^{N} n^2(k)}\right). \tag{5.4}$$

Das globale SNR wird über das gesamte Signal, daß heißt $k = 1, 2, \ldots, N$ bestimmt. Beim *segmentalen SNR* erfolgt dagegen die Berechnung über M kurze Signalsegmente der Länge K. Anschließend wird das Zeitmittel über alle M Signalabschnitte bestimmt:

$$SNR_{SEG} = \frac{1}{M}\sum_{m=1}^{M}\left[10\log\left(\frac{\sum_{k=1}^{K} s^2(k+mK)}{\sum_{k=1}^{K} n^2(k+mK)}\right)\right]. \tag{5.5}$$

Im Vergleich zum globalen SNR nach (5.4) reagiert dieses Maß empfindlicher auf Abschnitte geringer Signalaussteuerung bei konstanter Störleistung und weniger empfindlich auf Abschnitte verstärkter Störung bei konstantem Signalverhalten. Diese Signalabschnitte würden mit extrem großen negativen lokalen SNR-Werten zu stark in die Berechnung nach (5.5) eingehen und müssen daher gesondert betrachtet werden. Die Verbesserung des segmentalen Signal-Störabstandes $SNRI_{SEG}$ berechnet sich mit Gleichung (5.5) in Abschnitten mit Sprachaktivität zu:

$$SNRI_{SEG} = \frac{1}{M}\sum_{m=1}^{M}\left[10\log\left(\frac{\sum_{k=1}^{K} s^2(k+mK)}{\sum_{k=1}^{K} \hat{s}^2(k+mK)}\right)\right], \tag{5.6}$$

wobei das Sprachsignal $s(k)$ mit dem geräuschreduziertem Sprachsignal $\hat{s}(k)$ verglichen wird.

5.3.2 LPC- und kepstrale Distanz

Mit der Annahme, daß Verständlichkeit das Hauptziel und damit die Lautinformation wichtiger ist als die spektralen Feinstrukturen des Signals, kann man sich auf den Vergleich der spektralen Einhüllenden von entstörtem Nutzsignal und eigentlicher Störung zurückziehen. Weitgehende Unempfindlichkeit gegen Zeitverschiebungen und Amplitudenschwankungen zeigen parametrische Distanzmaße. Zweckmäßig ist es, parametrische Distanzmaße zu verwenden, die auf dem linearen Spracherzeugungsprozeß beruhen.

In den Abschnitten 6.4.2 und 6.4.3 werden die Verfahren der linearen Prädiktion und Kepstral-Analyse vorgestellt. Handelt es sich bei dem Originalsignal um Sprache, so eignen sich beide Verfahren besonders gut für einen Vergleich zwischen Originalsignal $s(k)$ und geräuschreduziertem Signal $\hat{s}(k)$ anhand der Koeffizienten $c_{LP}(k), \hat{c}_{LP}(k)$ und $c_{CC}(k), \hat{c}_{CC}(k)$ gemäß (6.46) und (6.53). Auf der Grundlage der LP- bzw. CC-Koeffizienten läßt sich damit ein LPC-Abstand Δ_{LPC} angeben:

$$\Delta_{LPC} = \frac{1}{M} \cdot \sum_{m=0}^{M-1} \left\{ \frac{\sum_{k=0}^{K-1} c_{LP}^2(k,m) \cdot \left| c_{LP}(k,m) - \hat{c}_{LP}(k,m) \right|}{\sum_{k=1}^{K} c_{LP}^2(k,m)} \right\}, \tag{5.7}$$

wobei *k* den Index der *K* Koeffizienten und *m* das *m*-te von insgesamt *M* Zeitfenstern bezeichnet. Für die kepstrale Distanz Δ_{CC} ergibt sich analog:

$$\Delta_{CC} = \frac{1}{M} \cdot \sum_{m=0}^{M-1} \left\{ \frac{\sum_{k=0}^{K-1} c_{CC}^2(k,m) \cdot \left| c_{CC}(k,m) - \hat{c}_{CC}(k,m) \right|}{\sum_{k=1}^{K} c_{CC}^2(m,k)} \right\}. \tag{5.8}$$

Die LPC- und die kepstrale Distanz werden in [dB] angegeben.

5.3.3 Bark-Distanz

Während in Abschnitt 5.3.2 Qualitätsmaße definiert wurden, die auf dem Spracherzeugungsprozeß basieren, wird nun eine Möglichkeit gezeigt, auch ein gehörrichtiges Abstandsmaß zu bestimmen. Zur instrumentellem Bewertung des Sprachsignals sind derartige Abstandsmaße am besten geeignet, da sie sich auf den beim Hörer wirklich einstellenden Höreindruck beziehen. Dazu werden die Intensitäten des Eingangsignals $I_s(z_k, m)$ und des Ausgangssignal $\hat{I}_s(z_k, m)$ auf der Bark-Skala z_k entsprechend (3.36) pro *m*-ten Analysefenster ins Verhältnis gesetzt und über *M* Zeitframes gemittelt. Das liefert die *Bark-Distanz* Δ_{Bark} mit:

$$\Delta_{Bark} = \frac{1}{M} \cdot \sum_{m=0}^{M-1} \left\{ \frac{\sum_{k=1}^{K} I_s^2(z_k,m) \cdot \left| I_s(z_k,m) - \hat{I}_s(z_k,m) \right|}{\sum_{k=1}^{K} I_s^2(z_k,m)} \right\}. \tag{5.9}$$

In Sprachpausen ist die Beurteilung der Sprachsignalverbesserung und Geräuschreduktion problematisch. Natürlich können dort keine Verzerrungen des Sprechersignals angegeben werden, so daß die Messung der LPC- und kepstralen Distanz auf stark fehlerhafte Werte führen würde. Dagegen kann die Dämpfung der Störanteile zwar gemessen werden, sie ist aber normalerweise deutlich höher als bei Sprachaktivität. Es hat sich gezeigt, daß vor allem der Charakter der Reststörungen in den Sprachpausen großen Einfluß auf die Hörempfindung hat. Treten bei-

spielsweise starke Störsignalverzerrungen (z.B. „musical tones“) auf, so wird dies als sehr störend empfunden. Um dennoch ein Vergleich und die Optimierung verschiedener Verfahren zu ermöglichen, wird die gehörrichtige Bark-Distanz gemäß Abschnitt 5.3.3 nur während Sprachaktivität berechnet.

In Sprachpausen wird dagegen zur Berechnung der Bark-Distanz die Intensität der ungedämpften Störung $I_n(z_k, m)$ mit der Intensität der resultierenden Reststörung $\hat{I}_n(z_k, m)$ verglichen. Demnach wird die Barkdistanz $\Delta_{Bark}^{(n)}$ in Sprechpausen wie folgt berechnet:

$$\Delta_{Bark}^{(n)} = \frac{1}{M} \cdot \sum_{m=0}^{M-1} \left\{ \frac{\sum_{k=1}^{K} I_n^2(z_k, m) \cdot \left| I_n(z_k, m) - \hat{I}_n(z_k, m) \right|}{\sum_{k=1}^{K} I_n^2(z_k, m)} \right\}. \tag{5.10}$$

5.3.4 Spektrogramm und Barkgramm

Die wohl bekannteste Form der spektralen Kurzzeitanalyse und „makroskopische“ Bewertung eines Signals stellt die Darstellung in einem *Spektrogramm* dar. Dabei wird das zu analysierende zeitdiskrete Signal $s(k)$ in einem zeitlich verschobenen Hamming- oder Hanningfenster $w(k)$ analysiert. Für jeden Signalabschnitt wird dann die Kurzzeitleistungsdichte ermittelt und wie in Abbildung 5.8 dargestellt. Wegen des Unschärfeprinzips zwischen Frequenz- und Zeitauflösung ergibt sich ein Kompromiß: Wird das Fenster $w(k)$ zu groß gewählt, gehen zeitliche Feinstrukturen von $s(k)$ verloren. Wählt man $w(k)$ zu klein, leidet die Frequenzauflösung oder der Einfluß der Fensterung verfälscht das Ergebnis.

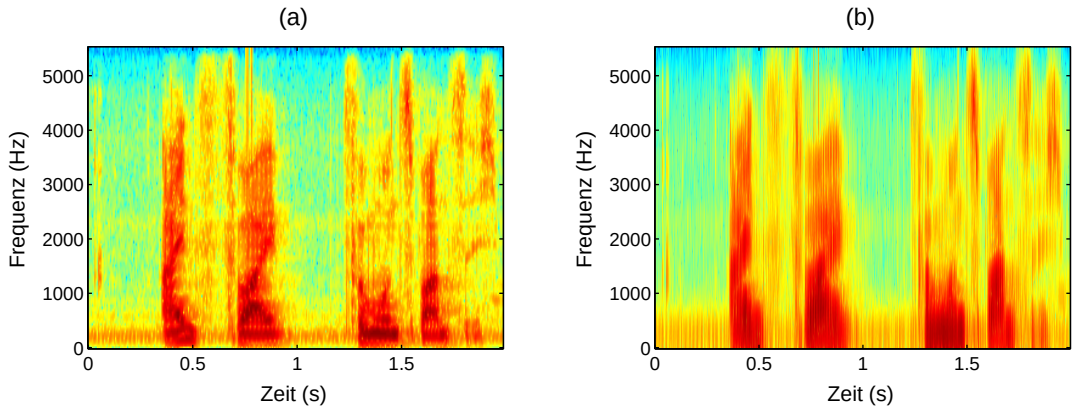


Abbildung 5.8 Dargestellt sind zwei Spektrogramme, die für dieselbe Äußerung im Fahrzeug aufgezeichnet und berechnet wurden. In **(a)** hat das Analysefenster eine Länge von 256 und in **(b)** eine Länge von 64 Abtastungen. Im direkten Vergleich sind deutlich die Unschärfen im Zeit- bzw. Frequenzbereich sichtbar. Rote Farbe: hohe Leistungsdichte, blau: niedrige Leistungsdichte.

Wird der Zeitverlauf der Intensitäten auf der Bark-Skala aufgetragen, so ergibt sich eine Darstellung, die als *Barkgramm* bezeichnet wird, siehe Abbildung 5.9. Das Barkgramm gibt den tatsächlichen Verlauf der Leistungsdichte auf der psychoakustischen Lautheitsskala wieder. Diese Darstellung ist an die Frequenzgruppenauflösung des menschlichen Gehörs angelehnt.

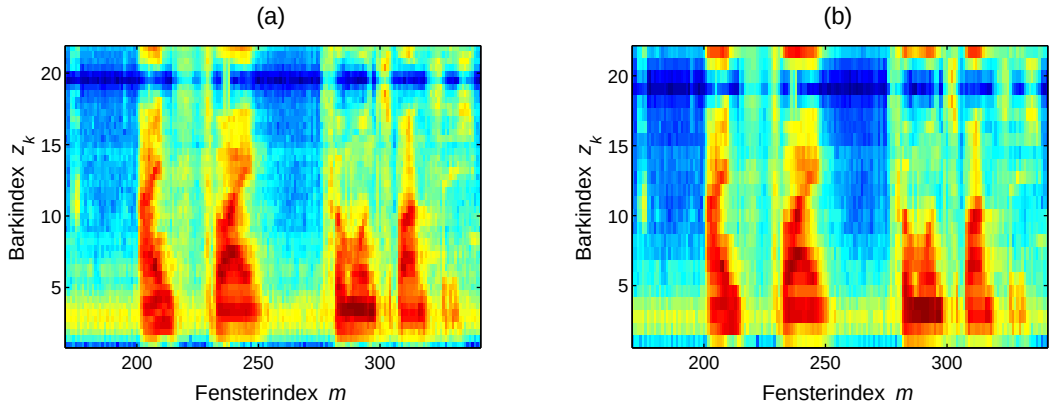


Abbildung 5.9 Barkgramm mit unterschiedlicher Auflösung. **(a)** 0.5 Bark **(b**) 1.0 Bark. Erklärung der Farben: Rot: hohe Leistungsdichte, blau: niedrige Leistungsdichte.

Eine einfache Möglichkeit das Spektrum eines zeitvarianten Signals analytisch zu beschreiben, stellt der Frequenzbandvektor $\mathbf{c}_{FB}(m)$ dar. Zur Bildung der B Frequenzbandkoeffizienten $c_{FB}^{(j)}(m)$, mit $\mathbf{c}_{FB}(m) = [c_{FB}^{(1)}(m), c_{FB}^{(j)}(m), \ldots, c_{FB}^{(B)}(m)]^{\mathrm{T}}$ und $1 \le j \le B$, wird das Kurzzeitleistungsdichtespektrum $|S(n, m)|^2$ in B Intervalle I_j eingeteilt. Da das LDS gerade und periodisch ist, brauchen nur die ersten $\frac{K}{2} + 1 = 129$ Werte des LDS betrachtet zu werden. In jedem Frequenzintervall I_j wird die mittlere Leistung des Signals durch die Mittelung der LDS-Werte bestimmt:

$$c_{FB}^{(j)}(m) = \frac{2}{K} \sum_{n \in I_j} |S(n,m)|^2. \tag{5.11}$$

Die Frequenzbandkoeffizienten $c_{FB}^{(j)}(m)$ bilden die einzelnen Komponenten des Merkmalsvektors $\mathbf{o}(m)$, wobei der Merkmalsverktor $\mathbf{o}(m)$ auf eins normiert wird, um von der Signalaussteuerung unabhängig zu werden. Mit $1 \le j \le B$ erhält man schließlich:

$$\mathbf{o}(m) = \frac{\mathbf{c}_{FB}(m)}{\sqrt{\sum_{j=1}^{B} \left| c_{FB}^{(j)}(m) \right|^2}}. \tag{5.12}$$

6 Psychoakustische Signalverbesserung

Da a priori kaum Kenntnisse von Sprachsignal, Störung und akustischer Umgebung vorliegen, sind Schätzfehler bei der Trennung der Signale und damit verbundene Verzerrungen des resultierenden Sprach- und Störsignals unvermeidbar. Dadurch führt eine vollständige Geräuschreduktion nicht zwangsläufig zu einer sofortigen Verbesserung der Sprachverständlichkeit. Durch die Nutzung psychoakustischer Maskierungseffekte werden hingegen derartige Verzerrungen und Restgeräusche verdeckt, sofern sie unterhalb der Mithörschwelle im Frequenz- und Zeitbereich liegen. Dementsprechend ist die Signalverarbeitung so zu optimieren, daß neben einer Reduktion der akustischen Störungen hörbare Verzerrungen des resultierenden Signals vermieden werden und sich eine bessere Verständlichkeit der Sprache im Vergleich zu herkömmlichen Reduktionsmethoden ergibt.

Die folgenden Untersuchungen und die psychoakustisch motivierten Modifikationen der Signalverarbeitung beziehen sich auf das Beispiel der einkanaligen Geräuschreduktion, sind aber genauso auf andere Verfahren anwendbar. Nachfolgend wird das für den Einsatz in Kraftfahrzeugen entwickelte Verfahren der psychoakustischen Geräuschreduktion vorgestellt. Dabei wird die Filterregel des linearen Reduktionsfilters (i.a. Wiener-Filter) so geändert, daß die Verzerrungen des Sprachsignals und die Reststörungen unterhalb der Maskierschwelle bleiben und damit verdeckt werden. Das vorgestellte Verfahren ist dabei nicht nur auf einkanalige Systeme beschränkt, sondern kann z.B. auch im Nachfilter einer Mehrkanallösung eingesetzt werden.

6.1 Grundstruktur und psychoakustische Modifikation

Das von Boll 1979 vorgestellte Verfahren der Spektralen Subtraktion [26] bildete die Ausgangsbasis für viele weiterführende Untersuchungen und Techniken der Geräuschreduktion. Darin wurden sowohl ein- als auch mehrkanalige Ansätze verfolgt, um den SNR-Gewinn zu erhöhen. Für viele Anwendungsfälle ist die Verbesserung des SNR nicht eigentliches Optimierungsziel. Erst die Erhöhung der Sprachverständlichkeit ergibt eine bessere Kommunikationsqualität, z.B. in Freisprecheinrichtungen für Telefone. Seit kurzem sind daher psychoakustische Methoden und Ansätze Gegenstand der Forschung. Die Vorteile, die durch

diese Methoden entstehen, sind offensichtlich: Nur in perzeptiven Bereichen des Nutzsignals oberhalb der psychoakustischen Maskierschwelle muß die Geräuschreduktion angewendet werden. Alle anderen Signalanteile sind ohnehin nicht wahrnehmbar. Das erhöht die Adaptionsgeschwindigkeit und Effektivität der Algorithmen. Außerdem können spektrale Schätzfehler durch psychoakustische Verdeckungseffekte so modifiziert werden, daß sie nicht mehr wahrnehmbar sind (noise suppression) oder zumindestens als weniger störend empfunden werden (noise shaping).

Bei der Spektralsubtraktion wird ausgenutzt, daß durch die Überlagerung von Nutzsignal $S(\Omega)$ und Störsignal $N(\Omega)$ die Leistungsdichte $R_{xx}(\Omega)$ des gestörten Signals $X(\Omega)$ gegenüber der Leistungsdichte des ungestörten Signals angehoben ist. Wird umgekehrt die Leistungsdichte der Störung von der Leistungsdichte des gestörten Eingangssignals subtrahiert, erhält man bei unkorrelierten Prozessen für Sprachsignal und Störung eine Schätzung der Leistungsdichte der ungestörten Sprache. Ausgehend von der allgemeinen Lösung des Optimalfilters gemäß Gleichung (2.99) für unkorrelierte Nutz- und Störsignale gelangt man durch Substitution von $z = e^{j\Omega}$ und mit $H(\Omega) = 1$ zum Frequenzgang des Optimalfilters $G(\Omega)$, wobei

$$G(\Omega) = \frac{R_{xx}(\Omega) - R_{nn}(\Omega)}{R_{xx}(\Omega)}. \tag{6.1}$$

Die Existenz einer solchen, durch ein kausales Digitalfilter realisierbaren Übertragungsfunktion ist keineswegs gesichert. Der Frequenzgang von $G(\Omega)$ ist wegen der Symmetrie der (Auto-)Leistungsdichtespektren reell und symmetrisch. Eine Übertragungsfunktion $G(\Omega)$, die derartige Eigenschaften besitzt, beschreibt ein nullphasiges Filter, d.h nach Rücktransformation in den Zeitbereich und entsprechender zeitlicher Verschiebung ergibt sich ein kausales, lineares Filter.

Die Herleitungen im Abschnitt 2.6 enthalten Idealisierungen, insbesondere die Annahme von Stationarität von Nutzsignal und Störung. Im Zeitbereich wurde deshalb von Erwartungswerten und im Frequenzbereich von Leistungsdichtespektren in allgemeiner Definition ausgegangen. Unter realen Bedingungen sind diese Annahmen nicht erfüllt, sondern es müssen die Kurzzeiteigenschaften der betreffenden Signale beachtet werden. Die Kurzzeiteigenschaften der Signale sind für einen endlichen Zeitraum bzw. für den m-ten endlichen Signalausschnitt, vgl. Abschnitt 2.3.1, zu bestimmen. Für die Lösung des Optimierungsproblems im Zeitbereich gemäß Abschnitt 2.6 bedeutet dies das Ersetzen der Korrelation in Gl. (2.98) durch Kurzzeit-Korrelationen. Für den Frequenzbereichsansatz (*Spektrale Subtraktion*) gemäß Beziehung

(2.99) sind die Kurzzeitleistungsdichtespektren mit endlich vielen Frequenzpunkten zu verwenden. Es liegt nahe, ein zeitbegrenztes Signalsstück $x(k, m)$, mit $k = 0, 1, \ldots, N-1$, unmittelbar einer DFT nach (2.18) zu unterziehen und das Kurz-Leistungsdichtespektrum im m-ten Signalrahmen durch Betragsquadrieren der Fouriertransformierten laut Gl. (2.40) zu schätzen. Insbesondere sind die Leistungsdichtespektren $R_{xx}(\Omega)$ und $R_{nn}(\Omega)$ durch die entsprechenden Kurzzeit-LDS $\tilde{R}_{xx}(n, m)$ und $\hat{R}_{nn}(n, m)$ gemäß Gl. (2.41) zu ersetzen. Die diskrete Fouriertransformierte $\hat{S}(n, m)$ der Nutzsignalschätzung $\hat{s}(k, m)$ erhält man mit Gleichung (2.99) zu:

$$\begin{aligned}\hat{S}(n,m) &= \frac{\tilde{R}_{xx}(n,m) - \hat{R}_{nn}(n,m)}{\tilde{R}_{xx}(n,m)} \cdot \big(S(n,m) + N(n,m)\big) \\ &= \left(1 - \frac{\hat{R}_{nn}(n,m)}{\tilde{R}_{xx}(n,m)}\right) \cdot X(n,m) \\ &= G_W(n,m) \cdot X(n,m).\end{aligned} \tag{6.2}$$

Das ist die Berechnungsvorschrift eines approximierten Wiener-Filters mit der Übertragungsfunktion $G_W(n, m)$. In der Literatur werden verschiedene Varianten der Spektralsubtraktion beschrieben, die sich vor allem in der Schätzung der Leistungsdichte des Störsignals unterscheiden. Auf einige Verfahren wird im Abschnitt 6.2 und Abschnitt 6.4 eingegangen.

Abbildung 6.1 zeigt das Prinzip der nach psychoakustischen Gesichtspunkten modifizierten spektralen Subtraktion in einem Blockschaltbild.

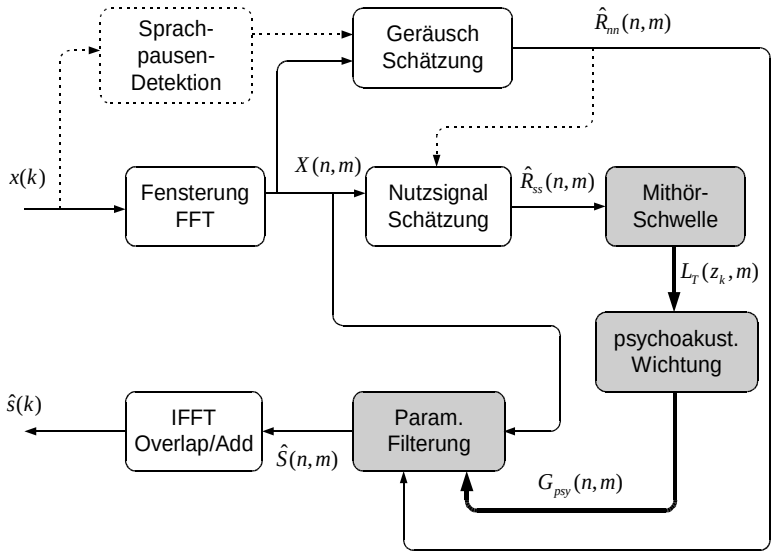


Abbildung 6.1 Blockdiagramm für das vorgestellte psychoakustische Geräuschreduktionsverfahren. Die psychoakustischen Modifikationen sind grau, optionale Komponenten sind gestrichelt dargestellt.

Das zeitdiskrete Eingangssignal $x(k)$ setzt sich aus dem Sprachsignal $s(k)$ und dem additiv überlagerten Störsignal $n(k)$ zusammen. Die Analyse und Filterung der einzelnen Signale in Abbildung 6.1 erfolgt im Frequenzbereich. Dazu wird das Eingangssignal $x(k)$ zunächst in $m = 1 \ldots M$ zeitliche Intervalle mit 50% Überlappung zerlegt. Dabei werden die Signalintervalle mit dem aus Abschnitt 2.3.1 bekannten Hanning-Fenster multipliziert. Durch abschnittsweise Fouriertransformation (WFFT) des Eingangssignals $x(k)$ ergibt sich das im allgemeinen komplexwertige diskrete Spektrum $X(n, m)$ gemäß (2.18).

Die modifizierte psychoakustische Geräuschreduktion findet im Frequenzbereich in drei Schritten statt: In der ersten Phase wird die spektrale Einhüllende des Störgeräusches $n(k)$ entsprechend Abschnitt 6.2 geschätzt. Daraus ergibt sich mit einfacher spektraler Subtraktion die Schätzung $\hat{R}_{ss}(n, m)$ für das LDS des unverfälschten Nutzsignals $s(k)$. Alternativ dazu ist die Schätzung des LDS $\hat{R}_{ss}(n, m)$ des Sprachsignals $s(k)$ anhand linearer Prädiktion möglich, siehe dazu Abschnitt 6.4.2. In Abbildung 6.1 wurde die Sprachpausendetektion gestrichelt eingezeichnet, da sie für das vorgestellte Verfahren explizit nicht notwendig ist. Die zweite Phase des Verfahrens ermittelt anhand der so gewonnenen Leistungsdichteschätzung $\hat{R}_{ss}(n, m)$ des Nutzsignals $s(k)$ die psychoakustische Mithörschwelle $L_T(z_k, m)$ und ermittelt dann die spektrale Gewichtsregel des psychoakustisch modifizierten Filters. In der dritten Phase erfolgt die eigentliche Filterung mit der parametrischen, psychoakustisch modifizierten Filterregel, siehe dazu Abschnitt 6.5.1. Am Ausgang des parametrischen Filters steht das geschätzte Spektrum $\hat{S}(n, m)$ des Sprachsignals $s(k)$ zur Verfügung. Nach abschnittsweiser Rücktransformation (IFFT) und Anwendung der *Overlap/Add*-Methode ergibt sich das zeitdiskrete geräuschreduzierte Sprachsignal $\hat{s}(k)$. Obwohl die Sprachsignalschätzung $\hat{s}(k)$ noch Restanteile der Störung $n(k)$ (spectral floor) enthält, werden alle Verzerrungen des Sprachsignals und der Reststörung unterhalb der Maskierschwelle gehalten. Sie werden verdeckt und sind damit nicht wahrnehmbar.

6.2 Schätzung der Störleistungsdichte

Für die Berechnung der Übertragungsfunktion des Spektralsubtraktionsfilters nach Gleichung (6.2) werden Schätzungen der Leistungsdichten des gestörten Sprachsignals und der Störung benötigt. Definitionsgemäß sind Leistungsdichten stochastischer Prozesse durch die Scharmittel ihrer Realisierungen bestimmt. Da Sprache und Störung keine ergodischen Prozesse darstellen, ist die Verwendung der Zeitmittel anstelle der Scharmittel der Realisierungen nicht zulässig. Das Störspektrum $R_{nn}(n, m)$ muß anhand einer Musterfunktion des stochastischen Prozesses geschätzt werden. Die Schätzung der Störleistungsdichte hat jedoch direkten Einfluß auf die Qualität der Geräuschreduktion. Fehlschätzungen von $\hat{R}_{nn}(n, m)$ bewirken sofort eine

Verschlechterung der Geräuschreduktion. Für das Wiener Filter nach (2.99) sind demnach das Leistungsdichtespektrum $\hat{R}_{xx}(n, m)$ des gestörten Nutzsignals und das Störleistungsdichtespektrum $\hat{R}_{nn}(n, m)$ zu schätzen. In einkanaligen Systemen können diese Größen nur indirekt aus dem gestörten Eingangssignal gewonnen werden, während bei Mehrkanallösungen das Autoleistungsdichtespektrum des ungestörten Nutzsignals direkt aus den Kreuzleistungsdichtespektren mehrerer Kanäle geschätzt wird. Dabei werden häufig Informationen bezüglich der räumlichen Ausbreitung der Sprache und Störung und der Mikrophonanordnung genutzt, siehe [24], [54], [113] und [201]. Diese Verfahren werden aber in Kraftfahrzeugen wegen der hohen Kosten und des eingeschränkten Bauraums nur selten angewendet. Einen günstigen Kompromiß zwischen Aufwand und qualitativer Verbesserung bieten zweikanalige Lösungen, wie z.B. in [113] vorgestellt. Im folgenden werden nur einkanalige Verfahren zur Schätzung der Stör- und Nutzsignalleistungsdichte näher beschrieben und dann ein neues Verfahren vorgestellt, das ohne explizite Pausenschätzung oder a priori Wissen der Signalstatistik auskommt. Als Grundvoraussetzung für die Funktion des Verfahrens werden drei schon bekannte Annahmen getroffen, die im allgemeinen für alle einkanaligen Schätzalgorithmen gemacht werden müssen. Sie werden hier noch einmal wiederholt:

- Nichtlineare Effekte durch die Generierung, Übertragung oder Wandlung der akustischen Wellen in elektrische Signale bzw. abgetastete Signale werden vernachlässigt. Das Eingangssignal $x(k)$ ist eine zeitdiskrete, lineare Funktion mit additiver Überlagerung von Sprachsignal $s(k)$ und Störung $n(k)$, wobei gilt $x(k) = s(k) + n(k)$.
- Sprachsignal $s(k)$ und Störung $n(k)$ sind unabhängige, mittelwertfreie stochastische Prozesse, die in begrenzten Zeitabschnitten stationär im weiteren Sinne sind. Stationarität im weiteren Sinne bedeutet, daß die ersten und zweiten statistischen Momente der Prozesse zeitinvariant sind.
- Die statistischen Eigenschaften des Störgeräusches $n(k)$ ändern sich bedeutend langsamer als die der Sprache $s(k)$.

6.2.1 Schätzung der Störleistungsdichte in Sprachpausen

Für stationäre Signale $x(k) = s(k) + n(k)$ ist die Schätzung des Störleistungsdichte einfach zu realisieren: Während einer Sprachpause liegt nur das Störsignal vor, dessen Leistungsdichte dann geschätzt werden kann. In folgenden Signalabschnitten wird davon ausgegangen, daß sich die statistischen Eigenschaften und das Leistungsdichtespektrum nicht ändern. Bei Stationarität oder zumindestens langsamer Zeitvarianz der Störung gilt das in Sprachpausen ermittelte Stör-LDS auch für spätere Zeitabschnitte mit Sprachaktivität. Um wirklich nur Abschnitte in Sprachpausen zu analysieren, wird eine *Sprachpausendetektion* benötigt. Diese Aufgabe ist

keineswegs trivial, da a priori wenig über den Charakter der Störung bekannt ist[1]. Zur einfachen Pausenerkennung eignet sich z.B. der *Voice-Activity-Detector* (VAD), der im GSM-System zur Senderabschaltung in Sprechpausen verwendet wird [51]. Selbst bei sehr langsam veränderlicher Störsignalcharakteristik sind nach einiger Zeit Sprachpausendetektion und Schätzung der Störleistungsdichte zu wiederholen. Bei größerem SNR verbessert sich die Schätzung der Störleistungsdichte. Die Fehler durch ungenaue Pausendetektion, die bei größerem SNR besser arbeitet, werden kleiner. Dennoch zeigt sich bei dynamischen, zeitvarianten Signalen ein schlechteres Schätzergebnis als bei nahezu stationären Störsignalen. Besonders deutlich ist dies beim Vergleich von konstanter Fahrt und Regenfahrt in Abbildung 6.2 zu erkennen.

Für den Vergleich mit anderen Schätzalgorithmen wurde ein Verfahren verwendet, das mit einer einfachen Schätzung der Sprachaktivität die Pausendetektion vornimmt. Dabei werden die Zeitfenster mit dem Index m als m_{aktiv} markiert, für die gilt:

$$P_x = \frac{1}{N}\sum_{n=1}^{N}\left[x(n)\right]^2 > D \cdot \sum_{k=-5}^{5} \tilde{R}_{xx}(n, m+k). \tag{6.3}$$

Mittels einfacher *moving-average*-Detektion nach Gleichung (6.3) werden die Sprachpausen gefunden. Dafür benötigt man eine feste Detektionsschwelle D mit $0 < D < 1$. Die Qualität der Pausendetektion hat großen Einfluß auf die Genauigkeit der Schätzung $\hat{R}_{nn}(n, m)$. Bei kleinerem SNR ist es schwieriger, die exakten Sprach-Pausen-Grenzen zu finden. Eine Fehldetektion bewirkt sofort eine Fehlschätzung der Störleistungsdichte. Die eigentliche Schätzung der Störleistungsdichte $\hat{R}_{nn}(n, m)$ erfolgt dann nach folgender Vorschrift:

$$\hat{R}_{nn}(n, m) = \begin{cases} \hat{R}_{nn}(n, m-1) & \text{für} \quad m = m_{aktiv} \\ \tilde{R}_{xx}(n, m) & \text{sonst,} \end{cases} \tag{6.4}$$

wobei die Kurzzeitleistungsdichte des gestörten Sprachsignals $x(k)$ mit $\tilde{R}_{xx}(n, m)$ bezeichnet ist. Ändert sich das in Sprachpausen geschätzte Störsignal in den nachfolgenden Sprachsignalabschnitten, so kommt es zu deutlichen Schätzfehlern, da die Schätzung des Störsignals in den Sprachpausen für die fortschreitende Zeit nicht mehr gilt. Auch der Charakter des Störsignals spielt eine große Rolle. Bei sehr dynamischen, sich mit der Zeit stark verändernden Geräuschen verschlechtert sich das Schätzergebnis, da auch innerhalb der Signalabschnitte nicht

[1] Ein besonders schwieriger Fall ergibt sich, wenn beispielsweise ein zweiter Sprecher das Störsignal $n(k)$ bildet oder wenn eine Pausendetektion in stark halliger Umgebung mit $\mathrm{H}(\Omega) \neq 1$ durchgeführt werden soll.

mehr von Stationarität ausgegangen werden kann. Während Sprachaktivität kommt es hierbei sofort zu fehlerhaften Schätzungen der Leistungsdichte der Störung. In [119] wird dieses Verfahren durch eine neuartige Methode der Störschätzung verbessert. Dabei wird zusätzlich untersucht, ob stimmhafte oder stimmlose Sprachabschnitte vorliegen. Abhängig davon werden tiefe oder hohe Frequenzen des Spektrums zur Schätzung der Störleistungsdichte herangezogen. Somit werden häufigere Aktualisierungen der geschätzten Störleistungsdichte möglich.

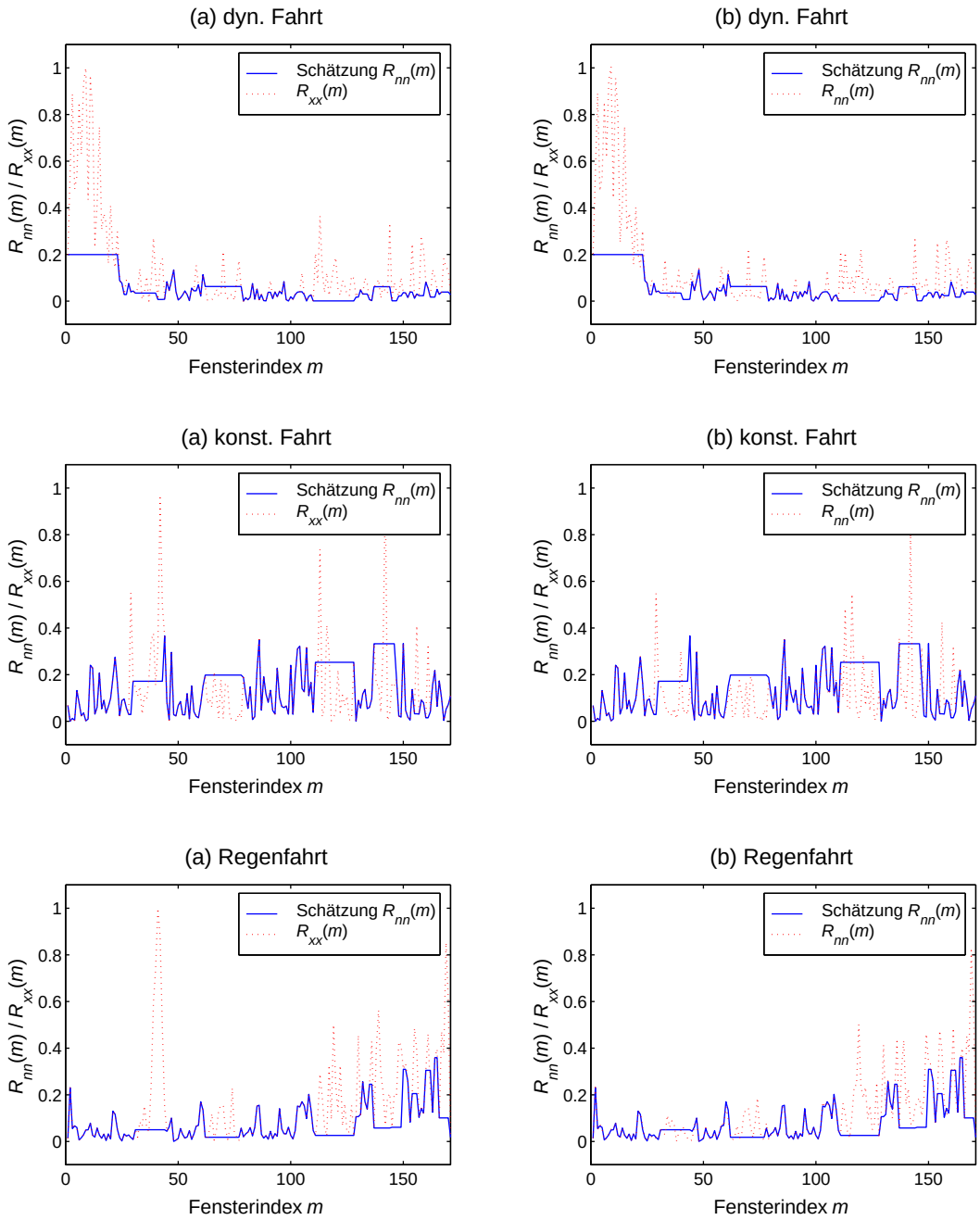


Abbildung 6.2 Schätzung der Störleistungsdichte in Sprachpausen. Aufnahme der Sprachprobe im BMW 528i touring bei unterschiedlicher Fahrweise. **(a)** zeigt LDS des gestörten Signals und der Störschätzung bei $f = 800$ Hz , **(b)** Vergleich LDS Störsignal mit LDS der Schätzung des Störsignals.

6.2.2 Spektraler Minimum-Schätzer

In letzter Zeit sind Verfahren vorgeschlagen worden, die ohne explizite Sprachpausendetektion auskommen. In [115] wurde eine derartige Approximation des Störleistungsdichtespektrums vorgestellt, die auf die Auswertung der statistischen Eigenschaften des Stör- und Sprachsignals beruht. Dabei wird in jedem *m*-ten Datenblock $x(k, m)$ die Kurzzeitleistungsdichte $\hat{R}_{xx}(n, m)$ geschätzt. Das innerhalb eines Intervalls von ca. 250 ms auftretende *Minimum* wird als Schätzwert für die aktuelle Störleistungsdichte $R_{nn}(n, m)$ verwendet. Die Zahl der für die Minimumsuche herangezogenen Stützstellen muß so groß sein, daß ein längerer Vokal nicht nach einiger Zeit unterdrückt wird. Andererseits muß sie klein gegenüber der Variabilität der Störung sein.

Das Verfahren macht explizit von der zeitveränderlichen Statistik von Sprach- und Störsignalen Gebrauch. Dabei wird angenommen, daß sich die Störung zeitlich langsamer ändert als das Sprachsignal.

Folgende Berechnungsvorschrift wird für den Schätzwert $\hat{R}_{nn}(n, m)$ der Störleistungsdichte genutzt, wobei die Rückgrifftiefe des Minimalfilters zu $p = 10...15$ gewählt wurde:

$$\hat{R}_{nn}(n,m) = \min\left(\hat{R}_{xx}(n,m), \hat{R}_{xx}(n,m-1), ..., \hat{R}_{xx}(n,m-p)\right). \tag{6.5}$$

Da das Minimum der Leistung des Signalgemisches stets kleiner ist als die tatsächliche Störleistung, ist der auf diese Weise ermittelte Schätzwert der Störleistungsdichte *nicht* erwartungstreu. Der Schätzwert $\hat{s}(n)$ der Größe $s(n)$ ist *erwartungstreu*, wenn der Schätzfehler im Mittel gleich Null ist. Es gilt dann nämlich:

$$E\{\hat{s}(n) - s(n)\} = 0. \tag{6.6}$$

In [115] kommt deshalb ein Korrekturfaktor, der in Abhängigkeit von der Signalstatistik berechnet wurde und den Algorithmus erwartungstreu macht, zum Einsatz. Gegenüber der Geräuschschätzung in Sprachpausen hat das Minimum-Verfahren den Vorteil, daß der Schätzwert der Störleistungsdichte auch während der Sprachaktivität aktualisiert werden kann. Das Verfahren liefert bei stationärer Störung geringfügig schlechtere Schätzergebnisse als ein Schätzverfahren mit optimalem Pausendetektor, ist diesem bei instationärer Störung jedoch überlegen: Durch die vergleichweise geringen Verzerrungen des Sprachsignals hinterläßt das Min-Verfahren beim Einsatz in Geräuschreduktionssystemen einen guten Höreindruck. Besonders bei kleinem SNR sind kaum Verfälschungen des Sprachsignals wahrzunehmen.

Abbildung 6.3 zeigt die Ergebnisse der Minimum-Schätzung der Störleistung bei $SNR = 5\,\text{dB}$ am Beispiel. Die Schätzung der Störleistungsdichte bildet einen spektralen „Teppich“ (*engl. spectral floor*), auf dem die Kurzzeit-Leistungsdichte $R_{xx}(n, m)$ des gestörten Sprachsignals aufsetzt. Die Rückgrifftiefe p gibt vor, wie schnell die Schätzung $\hat{R}_{nn}(n, m)$ dem Störsignal folgt.

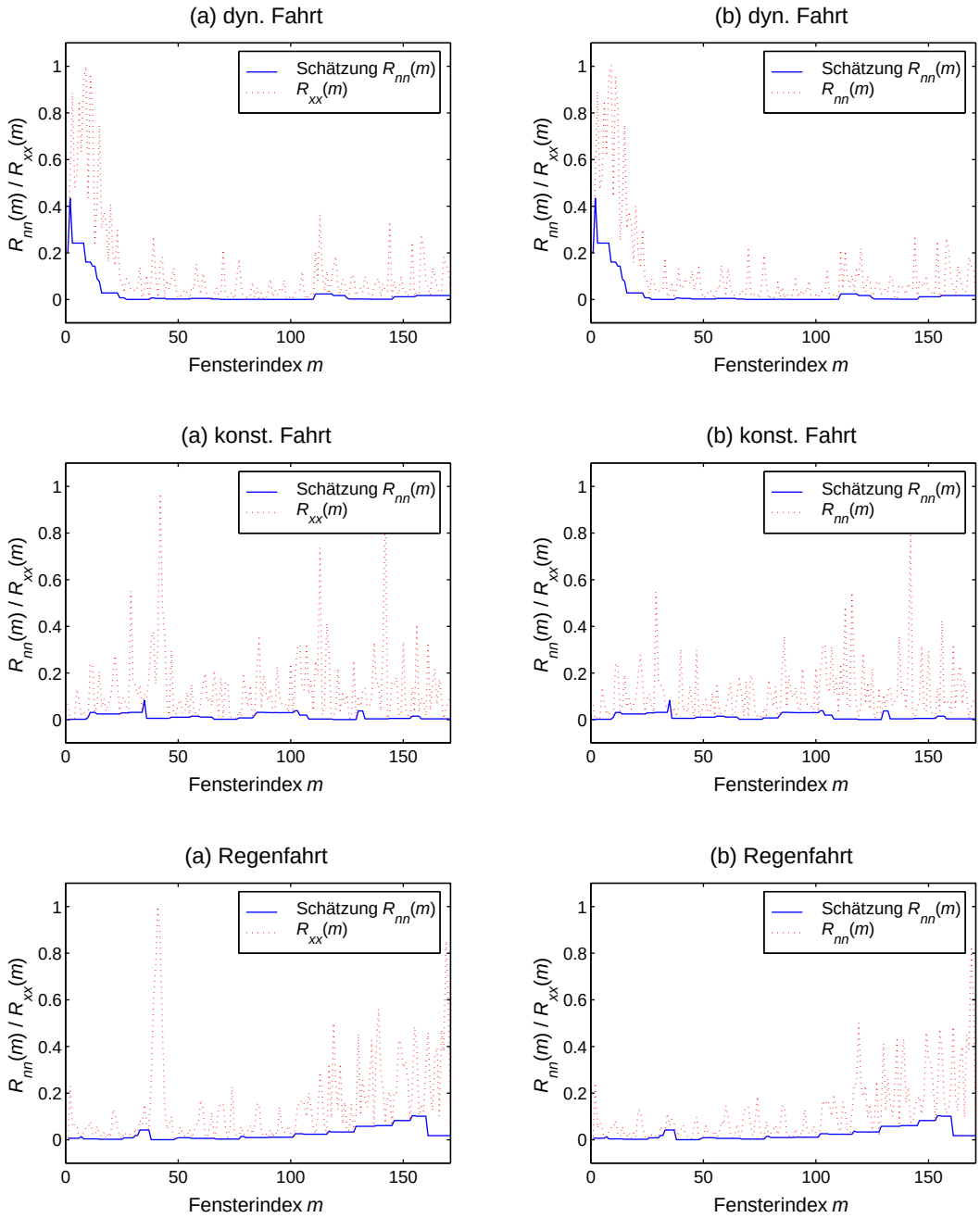


Abbildung 6.3 Schätzung des Störleistungsdichtespektrums durch Minimum-Verfahren bei $SNR = 5\,\text{dB}$. Aufnahme der gestörten Sprachprobe im BMW 528i touring bei unterschiedlicher Fahrweise bzw. Störsignalen gemäß Abschnitt 5.1.2. **(a)** zeigt das LDS des gestörten Signals und der dazugehörigen Störschätzung bei $f = 800\,\text{Hz}$ **(b)** Vergleich von LDS des Störsignals mit LDS der Schätzung des Störsignals.

6.2.3 CA-Verfahren zur Schätzung der Störleistungsdichte

In der Literatur sind weitere Schätzverfahren für spektrale Leistungsdichten, wie z.B. Eigenwertmethoden [73], modellbasierende Verfahren [33] oder lernbasierende Verfahren [94] bekannt geworden. Nachfolgend wird ein neues Verfahren zur Schätzung der Störleistungsdichte ohne explizite Sprachpausen- oder Sprachartdetektion, das hier so benannte *CA-Verfahren* (*Case-Approximations*-Verfahren) entwickelt und vorgestellt.

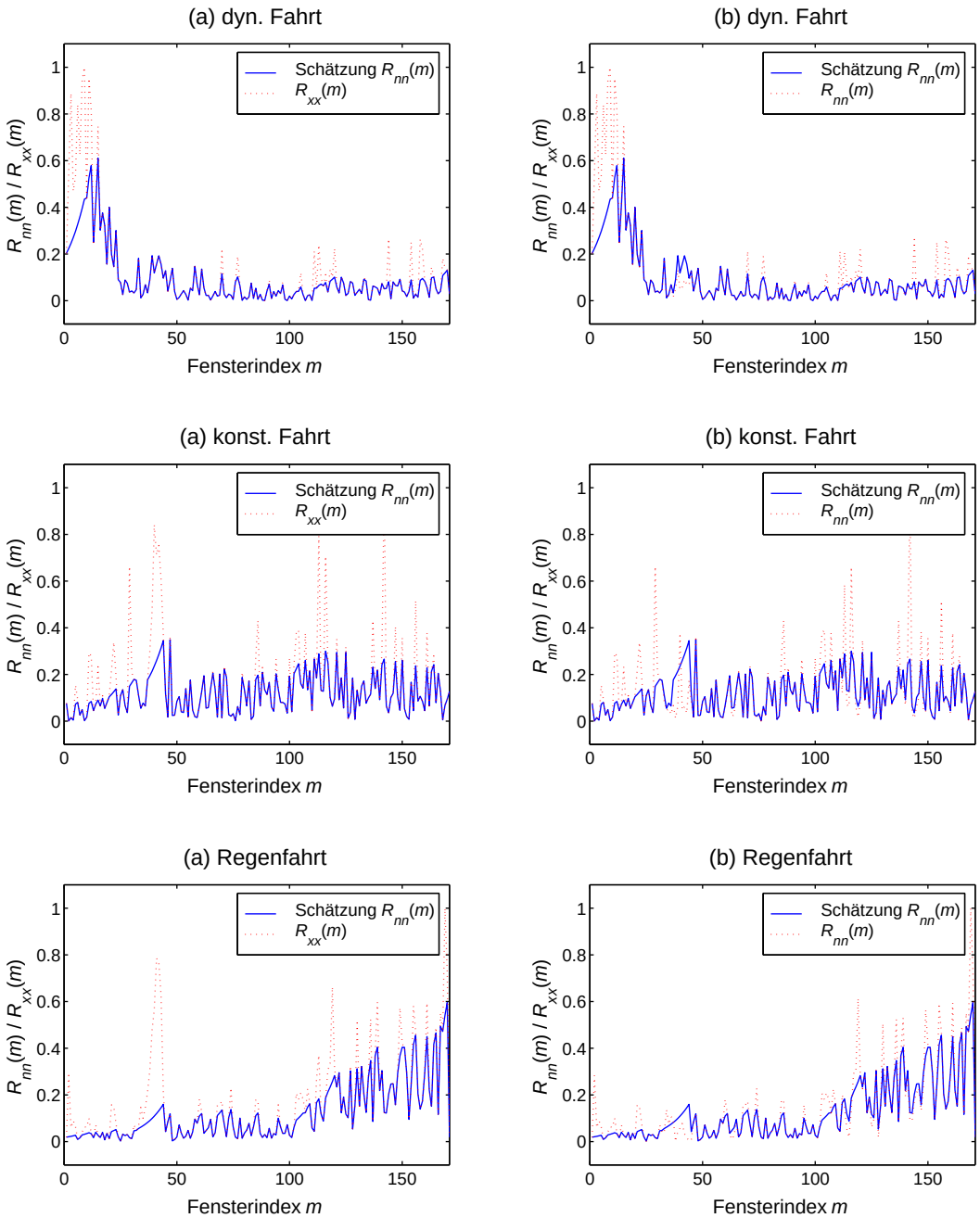


Abbildung 6.4 Schätzung des Störleistungsdichtespektrums mittels CA-Verfahren bei $SNR = 5$ dB. Aufnahme im BMW 528i touring gemäß Abschnitt 5.1.2. **(a)** LDS $\tilde{}$ des gestörten Signals $\hat{}$ und die dazugehörige Schätzung bei $f = 800$ Hz **(b)** Vergleich Kurzzeit-LDS $\tilde{R}_{nn}(m)$ mit Schätzung $\hat{R}_{nn}(m)$.

Je nach Verlauf (*case*) der Schätzung der Störleistungsdichte $\hat{R}_{nn}(n, m)$ im Vergleich zum Verlauf der Schätzung der Leistungsdichte des gestörten Eingangssignals $\hat{R}_{xx}(n, m)$ erfolgt die rekursive Schätzung der Störleistungsdichte durch *Approximation*. Ausgangspunkt der Überlegungen ist die Tatsache, daß sich in aufeinanderfolgenden Zeitrahmen $(m-1, m, m+1)$ die Leistungsdichte $R_{nn}(n, m)$ nur um einen endlich begrenzten Korrekturfaktor ändert (zeitliche Stetigkeitsbedingung physikalischer Signale). Berechnet man die Störleistungsdichte $\hat{R}_{nn}(n, m)$ im m-ten Zeitrahmen durch

$$\begin{aligned}\hat{R}_{nn}(n,m) &= \beta \cdot \hat{R}_{nn}(n,m-1) + (1-\beta) \cdot \left|X(n,m)\right|^2 \\ &= \beta \cdot \hat{R}_{nn}(n,m-1) + (1-\beta) \cdot \tilde{R}_{xx}(n,m),\end{aligned} \tag{6.7}$$

dann wird ein neuer Schätzwert $\hat{R}_{nn}(n, m)$ für die Störleistungsdichte aus dem vorhergehenden Schätzwert $\hat{R}_{nn}(n, m-1)$ gebildet und mit der aktuellen Kurzzeitleistungsdichte $\tilde{R}_{xx}(n, m)$ dem Spektralverlauf von $R_{xx}(n, m)$ nachgeführt. In Gleichung (6.7) kommt der Parameter β mit $0 < \beta < 1$ zum Einsatz, der das Tempo der Nachführung bestimmt. Durch z-Transformation von Gleichung (6.7) und Umformung erhält man die Übertragungsfunktion der Nachführung zu:

$$H_n(n,z) = \frac{\hat{R}_{nn}(n,z)}{\tilde{R}_{xx}(n,z)} = (1-\beta)\frac{z}{z-\beta} \cdot \tag{6.8}$$

Die Übertragungsfunktion des Schätzfilters erinnert an ein rekursives Tiefpaßfilter erster Ordnung. Das Filter wirkt für alle Frequenzstützpunkte gleichermaßen und unabhängig von n. Der Vorfaktor $1-\beta$ dient der Regelung der Gleichspannungsverstärkung des Filters auf eins. Abbildung 6.5 zeigt ein Schätzfilter, das mit der Übertragungsfunktion $H_n(n, z)$ nach Gleichung (6.8) die Störleistungsdichte $\hat{R}_{nn}(n, m)$ aus dem Kurzzeit-Leistungsdichtespektrum $\tilde{R}_{xx}(n, m)$ des gestörten Eingangssignals $x(k)$ schätzt.

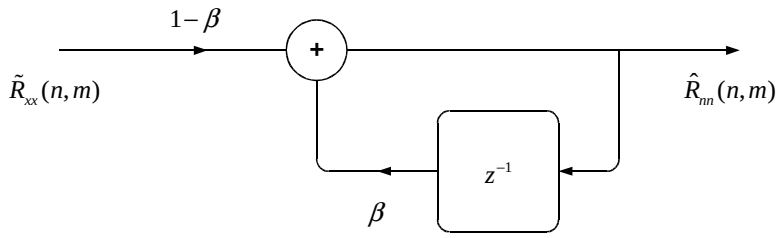


Abbildung 6.5 Filter zur Schätzung der Störleistungsdichte $\hat{R}_{nn}(n, m)$

Die Impulsantwort des Filters

$$h_n(n,m) = (1-\beta)\cdot\beta^m \quad \text{mit} \quad m \geq 0, \quad 0 < \beta < 1 \tag{6.9}$$

klingt mit β^m exponentiell ab. Dabei bestimmt der Parameter β mit $0 < \beta < 1$ die Abklinggeschwindigkeit über die m Zeitrahmen. Abbildung 6.6 verdeutlicht diesen Zusammenhang.

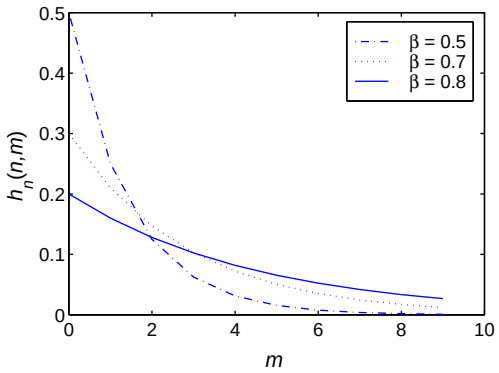


Abbildung 6.6 Darstellung des Abklingtempos in Abhängigkeit vom Parameter β

Für die Kurzzeit-Leistungsdichte $\tilde{R}_{xx}(n,m)$ des gestörten Sprachprozesses und die darin enthaltene geschätzte Störleistungsdichte $\hat{R}_{nn}(n,m)$ muß für alle Zeitrahmen m und alle Frequenzstützstellen n gelten:

$$\tilde{R}_{xx}(n,m) \geq \hat{R}_{nn}(n,m). \tag{6.10}$$

Um der Bedingung (6.10) zu genügen, wird eine Fallunterscheidung (*case*) durchgeführt, nach der das vorliegende Verfahren auch benannt wurde:

$$\hat{R}_{nn}(n,m) = \begin{cases} \beta\cdot\hat{R}_{nn}(n,m-1) + (1-\beta)\cdot\tilde{R}_{xx}(n,m), & \text{falls damit } \hat{R}_{nn}(n,m) < \tilde{R}_{xx}(n,m) \\ \tilde{R}_{xx}(n,m) & \text{sonst,} \end{cases} \tag{6.11}$$

wobei $0 < \beta < 1$. Die Schätzung der Störleistungsdichte $\hat{R}_{nn}(n,m)$ wird dem Verlauf des Kurzzeit-Leistungsdichtespektrums $\tilde{R}_{xx}(n,m)$ ab dem Zeitpunkt m nachgeführt, an dem das Kurzzeit-Leistungsdichtespektrum größer wird, als die Schätzung der Störleistungsdichte.

Der Parameter β bestimmt dabei die Geschwindigkeit der Nachführung. Er muß der Nutzsignalstatistik und Störsignalstatistik entsprechend angepaßt werden, siehe dazu Abschnitt 6.2.4.

Die Fallunterscheidung (6.11) bewirkt, daß bei stärker werdendem Sprachsignal, die Schätzung der Störsignalleistungsdichte gegen das Kurzzeit-Leistungsdichtespektrum des gestörten Sprachsignals konvergiert. Bei schwächer werdendem Sprachsignal, bis hin zur Sprechpause, geht Ungleichung (6.10) in eine Gleichung gemäß (6.11) über. Als Schätzung der Störleistungsdichte $\hat{R}_{nn}(n, m)$ wird in diesem Fall das aktuelle Kurzzeitleistungsdichtespektrum $\tilde{R}_{xx}(n, m)$ des gestörten Sprachsignals verwendet.

6.2.4 Bestimmung der Nachführungsgeschwindigkeit

Die Schätzgleichung (6.11) des CA-Verfahrens enthält den Parameter β, der für die Nachführungsgeschwindigkeit der Schätzung der Störleistungsdichte $R_{nn}(n, m)$ verantwortlich ist. Abbildung 6.7 zeigt die CA-Schätzung am Beispiel eines mit stationärem Fahrgeräusch gestörten Sprachsignals $s(k)$ für verschiedene β.

Für $0 \leq \beta \leq 1$ ergeben sich mit Gleichung (6.11) zwei triviale Fälle. Für $\beta = 0$ geht Gleichung (6.11) in $\hat{R}_{nn}(n, m) = \tilde{R}_{xx}(n, m)$ über. Die CA-Schätzung der Störung entspricht in diesem Fall dem Kurzzeit-Leistungsdichtespektrum des gestörten Eingangssignals $x(k)$.

Eine anschließende Spektralsubtraktion führt zur Totaldämpfung des gestörten Eingangssignals. Für $\beta = 1$ fällt die CA-Schätzung der Störleistungsdichte zunächst auf das Minimum des Verlaufs der Kurzzeit-Leistungsdichte $\tilde{R}_{xx}(n, m)$ des gemischten Eingangssignals $x(k)$ zurück. Danach wird die CA-Schätzung nicht mehr nachgeführt, alle Schätzwerte der Störleistungsdichte entsprechen dann dem Minimum des Verlaufs der Kurzzeit-Leistungsdichte $\tilde{R}_{xx}(n, m)$.

Die Bestimmung von β wird anhand des relativen Fehlers $|P_e(\beta)|$ gemäß Gleichung (6.12) durchgeführt. Dabei wird die absolute Schätzfehlerleistungsdichte $|\hat{R}_{nn}(n, m) - \tilde{R}_{nn}(n, m)|$, dem Betrag der Differenz zwischen dem geschätzten Leistungsdichtespektrum $\hat{R}_{nn}(n, m)$ und dem Kurzzeit-Leistungsdichtespektrum $\tilde{R}_{nn}(n, m)$ des Störers, verwendet.

Die absolute Schätzfehlerleistungsdichte $|\hat{R}_{nn}(n, m) - \tilde{R}_{nn}(n, m)|$ wird dann über alle Frequenzstützpunkte n, m aufsummiert und in Relation zur Kurzzeit-Leistungsdichte $\tilde{R}_{xx}(n, m)$ mit $\tilde{R}_{xx}(n, m) > 0$ gesetzt. Der relative Fehler $|P_e(\beta)|$ berechnet sich in Abhängigkeit von β zu:

$$|P_e(\beta)| = \frac{1}{M \cdot N} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \left| \frac{\hat{R}_{nn}(\beta, n, m) - \tilde{R}_{nn}(n, m)}{\tilde{R}_{xx}(n, m)} \right|. \tag{6.12}$$

Üblicherweise ist der Zugriff auf das Kurzzeitleistungsdichtespektrum des Störers innerhalb eines gestörten Nutzsignals nicht möglich. Um dennoch die Berechnung des relativen Fehlers $|P_e(\beta)|$ in Abhängigkeit von β durchzuführen, wird das gestörte Nutzsignal $x(k)$ durch manuelle Mischung aus reinem Sprachsignal $s(k)$ und separat aufgenommenem Störsignal $n(k)$ gemäß Abschnitt 5.1 gewonnen. Somit können sowohl die Art des Störgeräusches, wie auch das SNR frei gewählt werden.

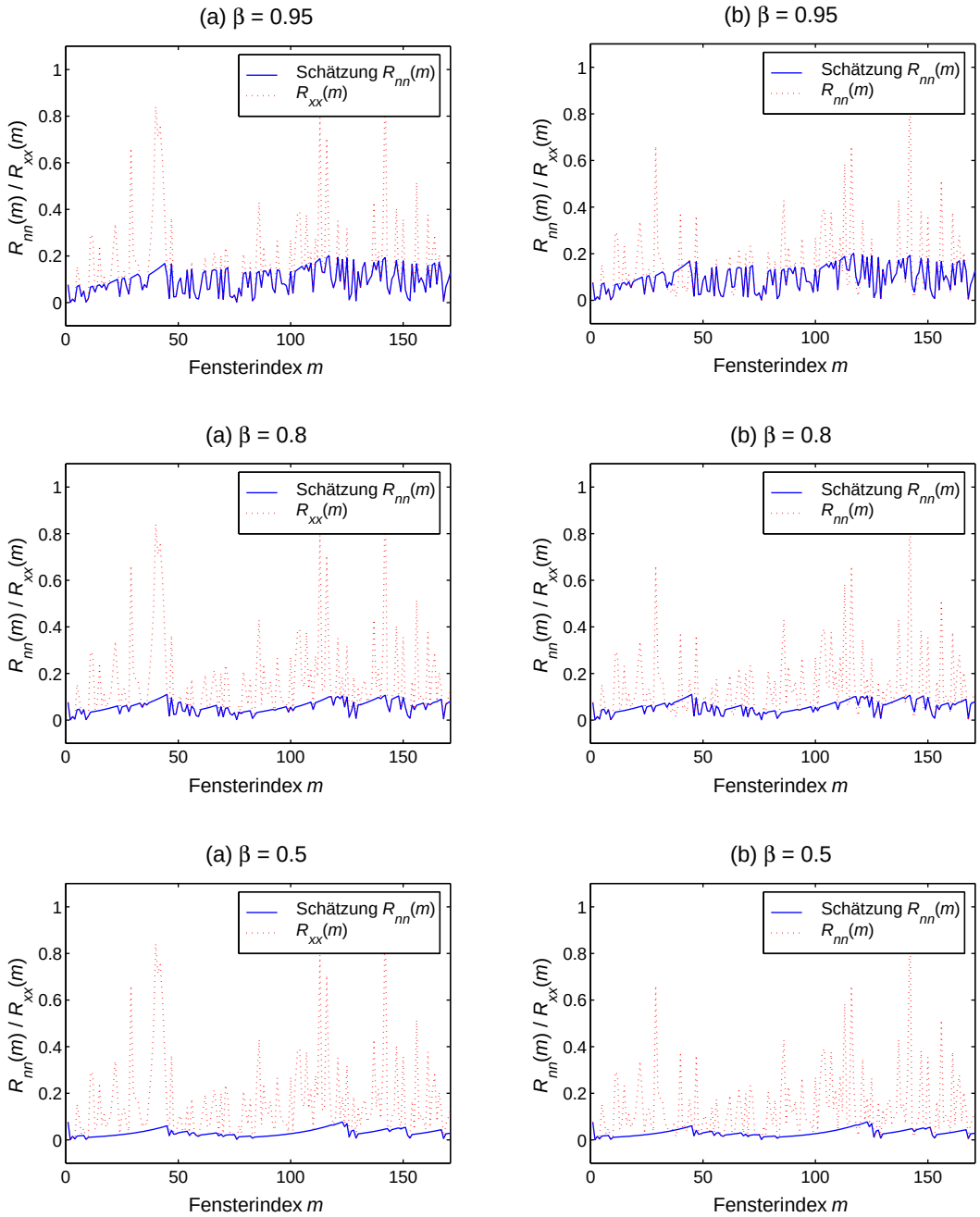


Abbildung 6.7 Schätzung des Störleistungsdichtespektrums mittels CA-Methode. **(a)** CA-Schätzung aus gestörtem Signal. **(b)** Vergleich der CA-Schätzung mit der Leistungsdichte des Störsignals (konstantes Fahrgeräusch mit SNR=5 dB . Aufnahme der gestörten Sprachprobe im BMW 528i touring).

Für die Bestimmung der „optimalen“ Nachführungsgeschwindigkeit β wurden unterschiedliche Sprech- und Störszenarien untersucht und die mittleren relativen Schätzfehler $|P_e(\beta)|$ nach Gleichung (6.12) berechnet. Dabei wurde β zwischen 0,6 und 1 variiert. Durch die Normierung der absoluten Fehlerleistung auf die Kurzzeit-Leistungsdichte $\tilde{R}_{xx}(n, m)$ wird $|P_e(\beta)|$ dimensionslos. Da die Schätzung $\hat{R}_{nn}(n, m)$ und die Kurzzeit-Leistungsdichte $\tilde{R}_{xx}(n, m)$ stets positiv sind und durch die Wahl des Verfahrens $0 \le \hat{R}_{nn}(n, m) \le \tilde{R}_{xx}(n, m)$ gilt, folgt:

$$0 \le |P_e(\beta)| \le 1. \tag{6.13}$$

Der relative Schätzfehler $|P_e(\beta)|$ ist für unterschiedliche Störszenarien in Abbildung 6.8 dargestellt. Dabei zeigen sich in den Verläufen des relativen Schätzfehlers $|P_e(\beta)|$ für alle Störarten lokale und globale Minima für $0.92 \le \beta \le 0.98$. Dies bestätigen auch Hörversuche, die unter Nutzung einer einfachen Wienerfilterung mit den geschätzten Störleistungsdichten ausgewertet wurden.

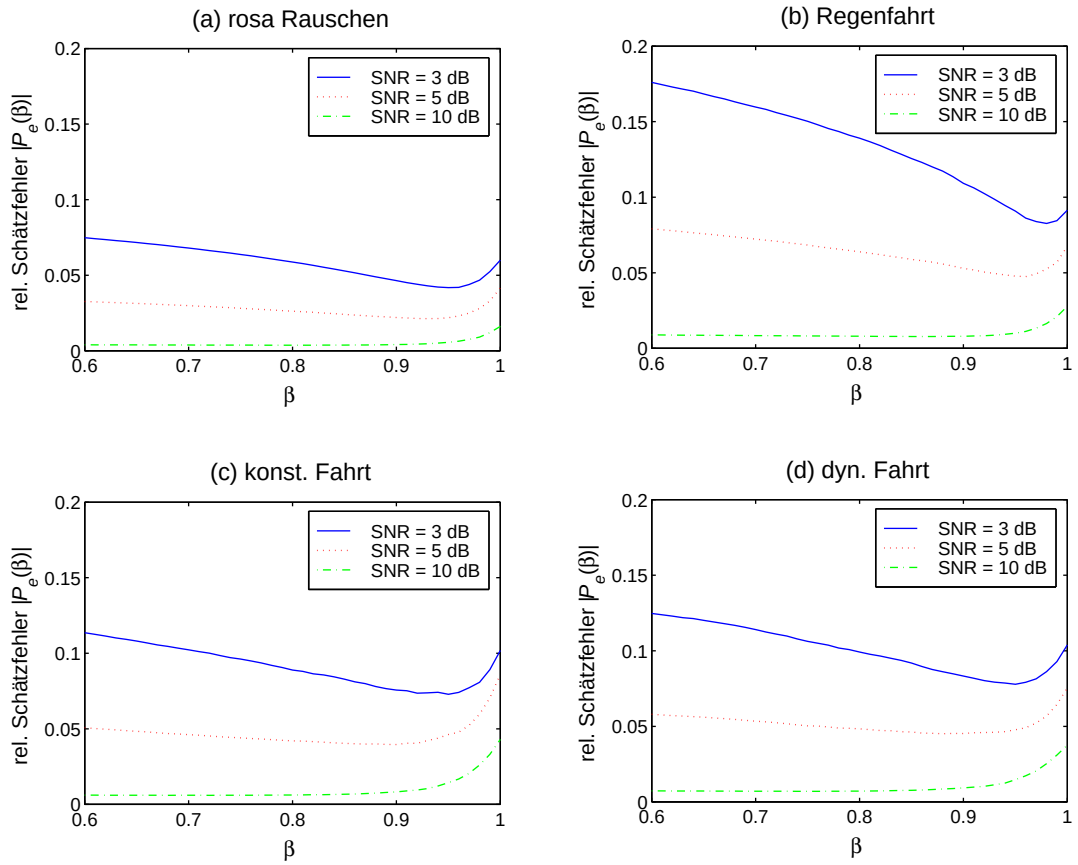


Abbildung 6.8 Untersuchung und Optimierung des Parameters β für die Schätzung der Störleistungsdichte mit dem CA-Verfahren. Aufnahmen im BMW 528i touring.

Besonders gute Ergebnisse bezüglich einer hohen Geräuschreduktion und vertretbaren Verzerrungen des Sprachsignals wurden mit $\beta = 0.95$ erzielt. Bemerkenswert ist, daß mit diesem Wert für unterschiedliche Störsituationen mit verschiedenen Geräuschen ähnliche Ergebnisse erreicht werden, siehe Abbildung 6.8.

Die Schätzung $\hat{R}_{nn}(n, m)$ ist genau dann erwartungstreu, wenn für den mittleren relativen Fehler $P_e(\beta)$ gilt:

$$P_e(\beta) = \frac{1}{M \cdot N} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \frac{\hat{R}_{nn}(\beta, n, m) - \tilde{R}_{nn}(n, m)}{\tilde{R}_{xx}(n, m)} = 0. \tag{6.14}$$

Dabei wurde die Fehlerleistungsdichte auf die Kurzzeit-Leistungsdichte $\tilde{R}_{xx}(n, m)$ normiert. So ist auch $P_e(\beta)$ dimensionslos.

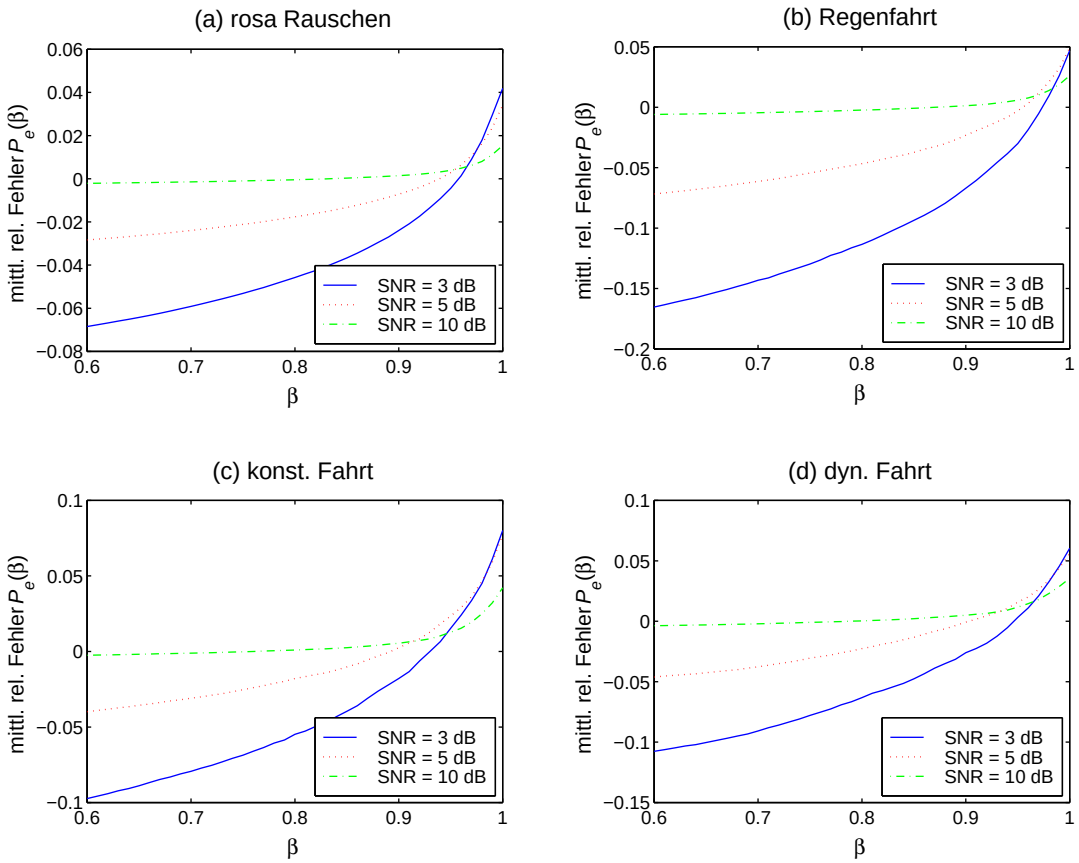


Abbildung 6.9 Untersuchung der Erwartungstreue der CA-Schätzung. Für $P_e(\beta) = 0$ ist die Schätzung der Störleistungsdichte $\hat{R}_{nn}(n, m)$ erwartungstreu. Folgende Geräusche wurden im BMW 528i touring für die Auswertung aufgezeichnet: **(a)** rosa Rauschen, **(b)** Regenfahrt, **(c)** gleichmäßige Fahrt und **(d)** dynamische Beschleunigungsfahrt.

Mit $0 \le \hat{R}_{nn}(n, m) \le \tilde{R}_{xx}(n, m)$ ergibt sich für den Wertebereich der mittleren relativen Störfehlerleistung:

$$-1 \le P_e(\beta) \le 1. \tag{6.15}$$

In Abbildung 6.9 ist der mittlere relative Fehler $P_e(\beta)$ für unterschiedliche Fahrsituationen dargestellt. Für verschiedene SNR und Geräusche ergeben sich in allen Diagrammen Schnittpunkte der Verläufe von $P_e(\beta)$ und der Nullordinate bei $\beta \approx 0.95$. Dies bestätigt die oben getroffenen Aussagen für den Wertebereich von β. Der Parameter kann mit

$$\beta = 0.95 \tag{6.16}$$

fest eingestellt werden.

6.2.5 Diskussion und Vergleich der Verfahren

Das neue CA-Verfahren benötigt entgegen der *Schätzung in Sprachpausen* aus Abschnitt 6.2.1 keine explizite Pausendetektion. Dadurch wird eine kritische Fehlerquelle ausgeschlossen. Mit dem CA-Verfahren erfolgt keine sogenannte Überschätzung, ein Effekt, der bei der Schätzung der Störleistungsdichte in Sprachpausen und der späteren spektralen Subtraktion negative Spektren erzeugt. Besser als bei der *Minimum-Schätzung* aus Abschnitt 6.2.2 treten mit dem CA-Verfahren im geschätzten Verlauf des Störleistungsdichtespektrums $\hat{R}_{nn}(n, m)$ keine Sprünge auf. Nur beim Übergang der Fallunterscheidung gemäß (6.11) ergeben sich Unstetigkeitsstellen, also genau dann, wenn der Verlauf der geschätzten Störleistungsdichte den Verlauf des Kurzzeit-Leistungsdichtespektrums des gestörten Sprachsignals schneidet.

In Abbildung 6.10 sind die absoluten Schätzfehler $|P_e(SNR)|$ für unterschiedlichen Störszenarien dargestellt. Das CA-Verfahren schneidet mit deutlich kleineren relativen Schätzfehlern in unterschiedlichen Fahr- und Geräuschsituationen ab. Dabei wird der resultierende Höreindruck im Einsatz zunächst nicht bewertet.

Das Ergebnis des Minimum-Verfahrens läßt sich auf einen womöglich unzureichend optimierten Parameter für die Einstellung der Erwartungstreue des Verfahrens zurückführen. Es konnte abschließend nicht geklärt werden, ob dafür eine adaptive Einstellung des Parameters geeigneter wäre. Für die nähere Behandlung dieser Thematik sei auf [115] verwiesen.

Zur Beurteilung der Schätzung der Störleistungsdichte mit dem Pausen-, Minimum- und CA-Verfahren wurden die Algorithmen in eine einfache Wienerfilterung integriert und die Ergeb-

nisse ohne weitere Nachbearbeitung begutachtet. Abbildung 6.11 zeigt die Resultate der mit unterschiedlichen Schätzverfahren durchgeführten Wienerfilterung anhand von Spektrogrammen der Eingangs- und Ausgangssignale des Wienerfilters.

Zunächst ist die deutlich stärkere Geräuschreduktion beim CA-Verfahren zu sehen. Dabei tauchen im Spektrogramm in Abbildung 6.11-f weniger Anteile des Sprachsignals auf, als bei den anderen beiden Verfahren. Dies könnte Indiz dafür sein, das zwar eine hohe SNR-Verbesserung erreicht wurde, aber dafür Verzerrungen des Sprachsignals in Kauf genommen werden müssen.

Die akustische Bewertung der Verfahren wurde nach Abschnitt 5 durchgeführt. Dabei kommt noch keine psychoakustische Nachverarbeitung zum Einsatz. Zunächst wird der sogenannte *Mean-Opinion-Score* bestimmt, eine subjektive Kenngröße für die akustische Wahrnehmung von Qualitätsunterschieden. Als Basis wurden jeweils vier unterschiedliche Geräuschszenarien mit einem Signal-Stör-Verhältnis von $SNR = 0$ dB ausgewertet.

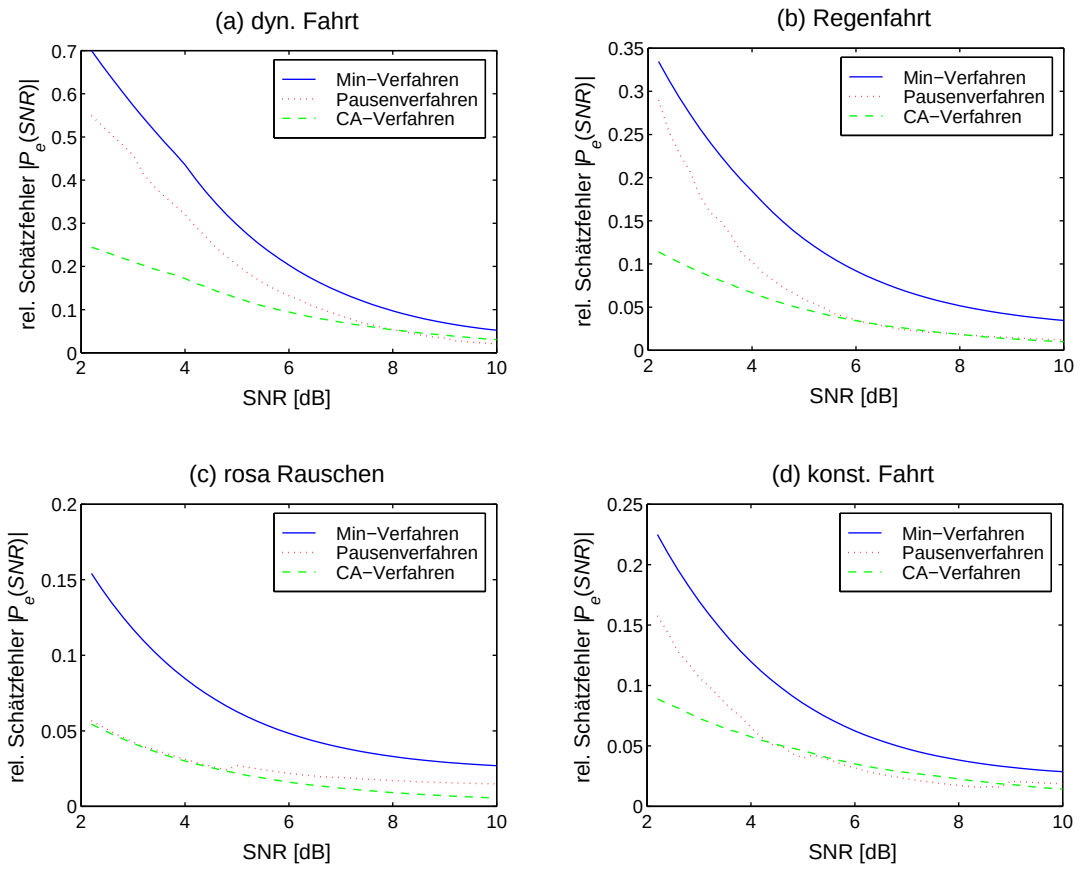


Abbildung 6.10 Berechnung der absoluten Fehlerleistung $|P_e(SNR)|$ gemäß Gleichung (6.12). Aufnahme der Geräusche im BMW 528i touring. **(a)** dynamische Beschleunigungsfahrt **(b)** Regenfahrt **(c)** rosa Rauschen **(d)** gleichmäßige Fahrt.

Tabelle 6.1 gibt in subjektiven Tests das wieder, was schon vorher der systematische Vergleich der Verfahren ergab. Prinzipiell hat das CA-Verfahren die höchste SNR-Verbesserung auf Kosten der stärkeren Verzerrung des Sprachsignals. Erstaunlich ist, daß das Minimum-Verfahren trotz geringster Geräuschreduktion den besten akustischen Eindruck hinterließ.

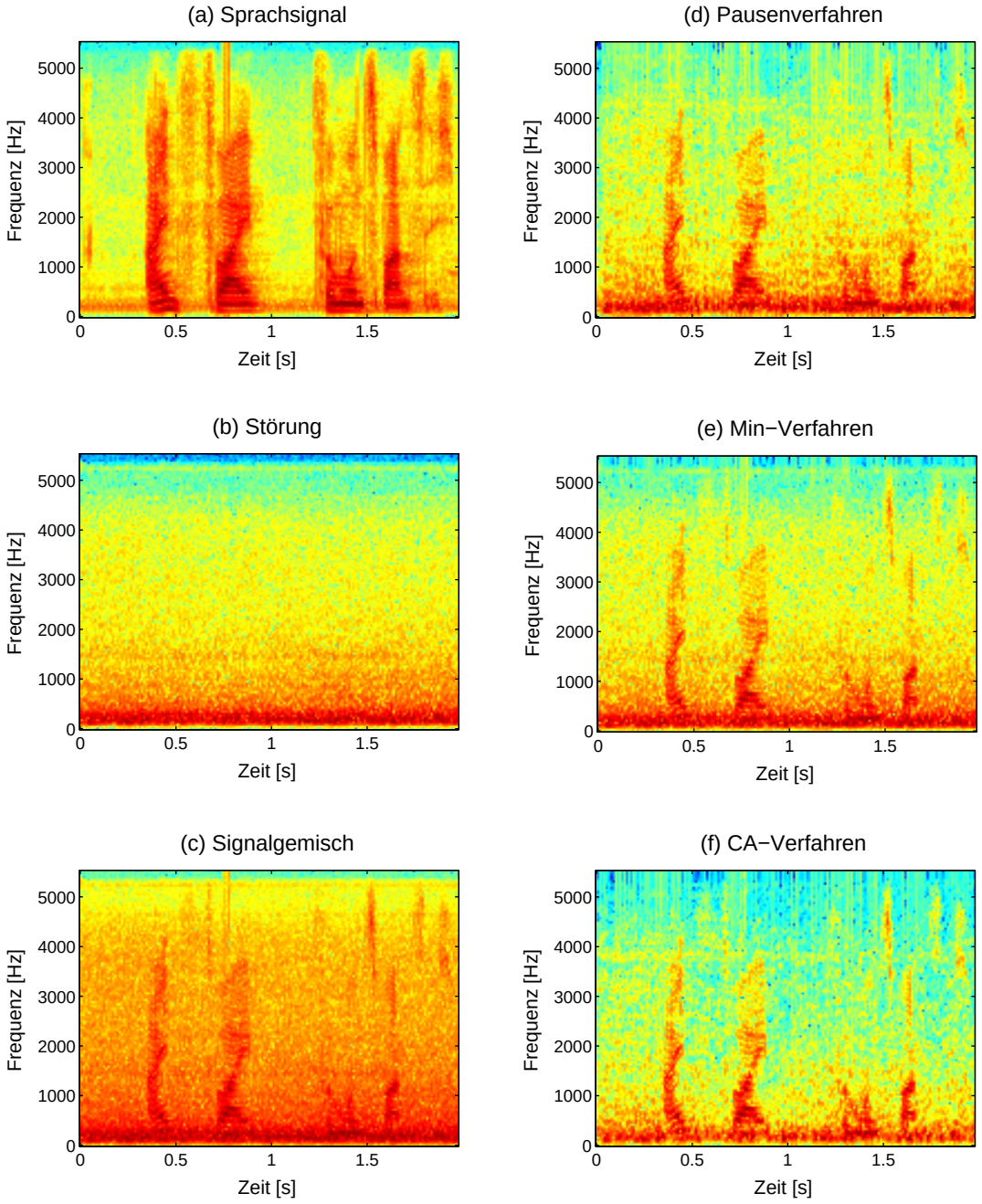


Abbildung 6.11 Darstellung der Spektrogramme der Eingangssignale und der mit einfacher Wienerfilterung und Schätzung der Störleistungsdichte mit dem Pausen-, Minimum- und CA-Verfahren gewonnenen geräuschreduzierten Signale $s_w(k)$. Aufnahme der Signale bei konstanter Fahrt im BMW 528i touring mit $SNR = 0$ dB.

Dies ist auf den vorsichtigen Ansatz zur Schätzung der Geräuschleistungsdichte zurückzuführen. Dabei wird praktisch nur der spectral floor entfernt, wodurch kaum Verzerrungen des Sprachsignals auftreten. Auf den instrumentellen Vergleich der drei Schätzverfahren wird an dieser Stelle verzichtet und auf die Gesamtbewertung psychoakustischer Geräuschreduktionsverfahren im Abschnitt 6.7 verwiesen.

Grundsätzlich motiviert bereits dieser Vergleich die Notwendigkeit der Optimierung der Geräuschreduktionsverfahren nach psychoakustischen Gesichtspunkten. Ziel ist eine effektive Geräuschreduktion mit geringer Verzerrung des Sprachsignals.

Verfahren	Verständlichkeit	Verzerrung des Sprachsignals	Charakter der Reststörungen	Reduktions-ergebnis
Eingangssignal				
rosa Rauschen	**3.3**	2.1		
konst. Fahrt	2.5	**1.5**	-	-
dyn. Fahrt	2.8	1.7		
Regenfahrt	3.0	1.7		
Pausenverfahren				
rosa Rauschen	2.8	3.2	3.0	
konst. Fahrt	2.5	2.9	**2.2**	2.9
dyn. Fahrt	3.0	3.1	2.7	
Regenfahrt	**3.3**	**3.**4	2.9	
Min-Verfahren				
rosa Rauschen	**3.0**	3.0	3.0	
konst. Fahrt	2.5	**2.7**	**2.2**	2.8
dyn. Fahrt	2.8	2.9	2.7	
Regenfahrt	2.9	3.3	2.9	
CA-Verfahren				
rosa Rauschen	2.9	3.7	2.6	
konst. Fahrt	2.5	2.8	**2.2**	2.9
dyn. Fahrt	3.1	3.2	2.5	
Regenfahrt	**3.2**	3.9	2.4	

Tabelle 6.1 *Mean-Opinion-Score* Bewertung gemäß Abschnitt 5.2.1. Besondere Auffälligkeiten wurden fett dargestellt. Das Reduktionsergebnis ergibt die gewichtete Gesamtnote aus den einzelnen gewichteten Teilbewertungen.

6.3 Bestimmung der globalen Mithörschwelle

Für die psychoakustische Modifikation einer herkömmlichen Geräuschreduktion ist die Bestimmung der globalen Mithörschwelle notwendig. Alle Verzerrungen oder Reststörungen unterhalb dieser Schwelle bleiben dem Hörer verborgen. Somit sind auch Schätzfehler der Geräuschleistungsdichte, sofern Sie unterhalb der Mithörschwelle bleiben, nicht mehr wahr-

nehmbar. Die Maskierschwelle wird bei Simultanverdeckung im m-ten Zeitrahmen bestimmt. Temporale Maskiereffekte werden vernachlässigt. Die Berechnung der psychoakustischen Maskierschwellen erfolgt mit mathematischen Modellen, die mit experimentellen Messungen an Probanden [202] und anschließender Approximation und Modellierung des Verlaufs der Maskierschwellen bestätigt wurden. Die in Abschnitt 3.3 beschriebenen Untersuchungen gelten für spezielle Kombinationen von Maskierer und Testsignal. Dabei sind Maskiereffekte für sinusförmige Maskierer mit frequenzgruppenbreitem, rauschartigem Testsignal besonders deutlich zu messen. Für derartige einfache Signalkonfigurationen wurden bereits genaue mathematische Beschreibungen der sogenannten *lokalen* Maskierschwellen z.B. in [202], [203], [180] und [172] gefunden.

Liegen dagegen komplexe Maskierer oder Testsignale vor, ist die experimentelle Messung und mathematische Formulierung von Maskiereffekten viel schwieriger. In diesem Fall sind die gefundenen psychoakustischen Modelle nur dann anwendbar, wenn der vorliegende komplexe Schall durch eine Kombination von einfachen Maskierern mit lokalen Maskierschwellen gebildet werden kann. Die resultierende *globale* Maskierschwelle ergibt sich durch *Superposition* der einzelnen lokalen Mithörschwellen. Geht man von einem linearen Verhalten der Hörempfindung aus, so führt das Superpositionsprinzip zur Bestimmung der globalen Maskierschwelle auf eine additive Überlagerung der einzelnen lokalen Maskierschwellen. Dieses Modell wird beispielsweise im ISO MPEG Standard [87] für die psychoakustische Datenreduktion verwendet. Dabei wird das Quellensignal in Bark-Subbändern z_k analysiert und in tonale (sinusartige) und nicht tonale (rauschartige) Komponenten zerlegt. Für die einzelnen tonalen und nicht tonalen Maskierer werden die lokalen Maskierschwellen berechnet und anschließend additiv überlagert. Damit ergibt sich die resultierende globale Maskierschwelle, siehe Abbildung 6.12.

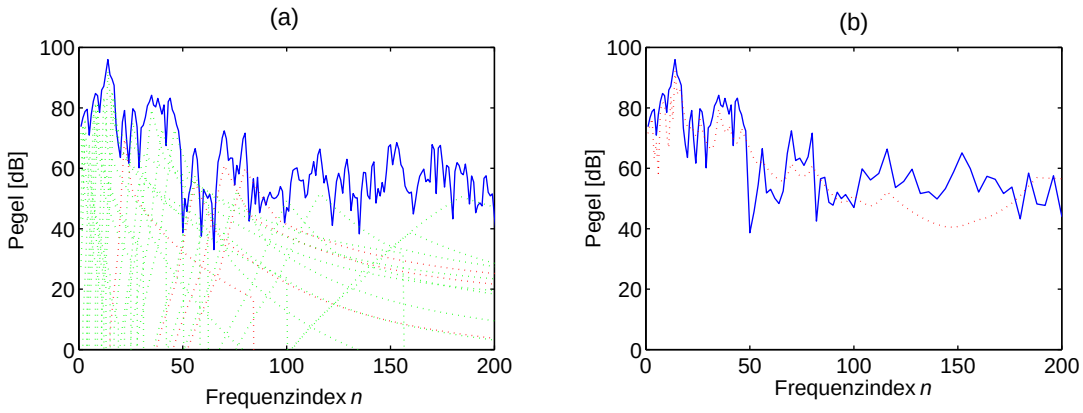


Abbildung 6.12 ISO MPEG Modell nach [87]. Bestimmung der psychoakustischen Maskierschwelle durch additive Überlagerung der lokalen Maskierschwellen. **(a)** Addition der lokalen Maskierschwellen ergibt **(b)** globale Maskierschwelle als gepunktete Linie.

Untersuchungen in [67] und [109] haben gezeigt, daß sich durch die additive Überlagerung der lokalen Maskierschwellen eine Diskrepanz zwischen der berechneten globalen Maskierschwelle und der wirklich meßbaren Mithörschelle ergibt. Die meßbare Mithörschwelle liegt viel höher als die durch additive Überlagerung berechnete globale Maskierschwelle. Zur Verbesserung wird ein nichtlineares Modell vorgestellt, wie es in ähnlicher Form in [110] eingeführt wurde. Dabei kommt den Aussagen aus Abschnitt 3.3.3 für die psychoakustische Modellierung komplexer Töne besondere Bedeutung zu.

6.3.1 Bestimmung der psychoakustischen Intensität des Sprachsignals

Sei $s(k)$ das abgetastete Nutzsignal mit der Schätzung der Kurzzeitleistungsdichte $\hat{R}_{ss}(n, m)$. Durch die nichtlineare Abbildung der Leistungsdichte des Nutzsignals $s(k)$ auf die Bark-Skala z_k entsprechend Abbildung 3.8 werden additive Maskiereffekte zwischen verschiedenen Frequenzgruppen ausgeschlossen. Die Intensität des Maskierschalls $I_M(z_k, m)$ in einem Bark-Subband z_k und dem *m*-ten Zeitfenster berechnet sich zu:

$$I_M(z_k, m) = \frac{1}{b_{k+1} - b_k} \sum_{n=b_k}^{b_{k+1}-1} \hat{R}_{ss}(n, m). \tag{6.17}$$

Mit der nichtlinearen Abbildungsfunktion $\gamma(z_k)$ gemäß (3.36) und

$$\Delta f = \frac{f_a}{K}, \tag{6.18}$$

wobei f_a die Abtastrate und K die Länge des DFT-Fensters darstellen, ergeben sich die von der Tonheit z_k abhängigen ganzzahligen Summationsgrenzen in (6.17) zu:

$$b_k = \text{int}\left(\frac{\gamma\left(z_k - \frac{1}{2}\Delta z \right)}{\Delta f} \right). \tag{6.19}$$

Damit ist die Intensität des Maskierschalls $I_M(z_k, m)$ von der Breite der kritischen Frequenzgruppe und der Auflösung der Analyse Δz abhängig.

Für z_k ergibt sich mit $0 < z_k \leq 24$:

$$z_k = k \cdot \Delta z. \tag{6.20}$$

Wird die Auflösung dieser Analyse erhöht, so erhöht sich auch die Anzahl der lokalen Maskierer. Die Schrittweite Δz sollte dennoch nicht zu klein gewählt werden. In diesem Fall reicht die Auflösung der Leistungsdichte gemäß Gleichung (6.18) nicht aus, um eine vollständige Abbildung auf die Tonheitsskala zu erreichen. Die optimale Schrittweite Δz ergibt sich aus dem Grenzfall $b_k - b_{k+1} \geq 1$, womit die Summation in Gleichung (6.17) gerade über eine Frequenzstützstelle n erfolgt. Für $f_a = 11025$ Hz und $N = 256$ erhält man $\Delta z = 0{,}43$.

6.3.2 Gehörrichtige Vorfilterung und Normierung

Die berechneten Intensitäten des Maskierschalls $I_M(z_k, m)$ werden als lokale Maskierer interpretiert. Die empfundene Lautheit dieser Maskierer ist frequenzabhängig. Die Kurven gleicher Lautheit $\Xi(z_k)$ aus Abschnitt 3.2.2.2 sind in Abbildung 3.9 dargestellt. Um der psychoakustischen Frequenzabhängigkeit zu entsprechen, werden die ermittelten Intensitäten $I_M(z_k, m)$ des Maskierschalls einer Vorfilterung unterzogen, so daß gilt:

$$L_M(z_k, m) = 10 \cdot \log[I_M(z_k, m) \cdot \Xi(z_k)]. \tag{6.21}$$

Die lokalen Maskierer $L_M(z_k, m)$ stellen damit die gehörrichtigen Komponenten des Nutzsignals $s(k)$ dar. Die lokalen Maskierer sind Ausgangspunkt für die Bestimmung der lokalen psychoakustischen Maskierschwellen $L_{T,i}(z_k, m)$. Dieses Konzept der Vorfilterung nimmt Rücksicht auf die Maskierung des Schalls in unterschiedlichen Frequenzgruppen.

6.3.3 Bestimmung der lokalen Maskierschwellen

Die aus Abschnitt 3.3.1 bekannten lokalen Maskierschwellen werden auf die lokalen Maskierer $L_M(z_k, m)$ angewendet. Diese lokalen Maskierschwellen, die grundsätzlich nur für die Anregung mit reinen Tönen gelten, können als Spreizfunktion interpretiert werden. Wie in Abbildung 6.13 gezeigt, wird die Spreizfunktion $L_{T,i}(z_k, m)$ vereinfachend durch die drei Parameter a_v, s_l und s_u bestimmt. Der Pegeloffset a_v zwischen dem lokalen Maskierer $L_M(z_i, m)$ und der Spitze der Spreizungsfunktion $L_{T,i}(z_k, m)$ beträgt ca. 12 dB, siehe Abbildung 3.13 und Abschnitt 3.3.1.3. Während die untere Flankensteilheit s_l der Spreizfunktion $L_{T,i}(z_k, m)$ nahezu pegelunabhängig mit

$$s_l = 17 dB / Bark \tag{6.22}$$

ist, ändert sich die obere Flankensteilheit s_u mit dem Pegel des lokalen Maskierers $L_M(z_k, m)$ nach folgender Eigenschaft, siehe auch [87]:

$$s_u = \frac{(22dB - 0.2L_M)}{Bark}. \tag{6.23}$$

Nach Abbildung 6.13 läßt sich die mathematische Repräsentation der *i*-ten Spreizungsfunktion $L_{T,i}(z_k, m)$ für lokalen Maskierer $L_M(z_i, m)$ an der Stelle z_i angeben:

$$L_{T,i}(z_k, m) = \begin{cases} L_M(z_i, m) - a_v - s_l \cdot (z_i - z_k); & z_k < z_i \\ L_M(z_i, m) - a_v - s_u \cdot (z_k - z_i); & z_k \geq z_i. \end{cases} \tag{6.24}$$

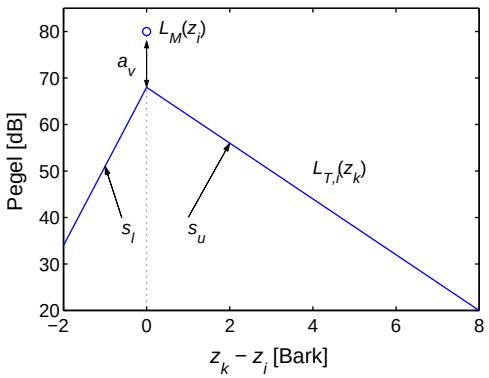


Abbildung 6.13 Lokale Maskierschwelle als Spreizfunktion. Die Spreizfunktion wird durch die beiden Geradenanstiege s_l und s_u sowie durch die Dämpfung a_v bestimmt.

6.3.4 Nichtlineare Superposition

Die Berechnung der globalen Maskierschwelle erfolgt durch Superposition der einzelnen lokalen Maskierer nach [110]. Die spektralen und temporalen Bedingungen und Effekte für die Anwendung des nichtlinearen Superpositionsprinzips bei simultaner Verdeckung wurden ausführlich in [82], [83] und [84] untersucht. Im Gegensatz zur linearen Addition der lokalen Verdeckungsschwellen nach [87] erfolgt die Superposition mit exponentieller Kompression nach folgendem Prinzip:

$$I_T(z_k, m) = \left[\sum_i I_{T,i}(z_k, m)^a \right]^{1/a}, \tag{6.25}$$

wobei

$$I_{T,i}(z_k, m) = 10^{L_{T,i}(z_k, m)/10}. \tag{6.26}$$

Die nichtlineare Addition wird auf die lokalen Intensitäten angewendet. Damit ergibt sich die globale Intensität des komplexen Maskiererschalls $I_T(z_k, m)$ in Abhängigkeit vom Parameter a. Für $a = 1$ geht Gleichung (6.25) in eine lineare Addition der einzelnen lokalen Maskierintensitäten über. Abbildung 6.14 zeigt die nichtlineare Superposition zweier lokaler, simultaner Maskierer mit den Verdeckungsschwellen $L_{T,1}(z_k, m)$ und $L_{T,2}(z_k, m)$. Durch die Superposition der i Maskierer nach (6.26) ergibt sich eine zusätzliche Maskierung $\Delta L_T(z_k, m)$ im Vergleich mit den einzelnen lokalen Mithörschwellen. Es gilt:

$$\Delta L_T(z_k, m) = L_T(z_k, m) - \max_i [L_{T,i}(z_k)]. \tag{6.27}$$

Die zusätzliche Maskierung beachtet damit die Unterschiede, die sich bei der Verdeckung durch komplexe Signale im Vergleich zu einfachen sinusartigen Signalen ergeben.

Wird die Auflösung der Analyse Δz lt. Gleichung (6.20) erhöht, so erhöht sich ebenfalls die Anzahl der lokalen Maskierer. Dadurch ergibt sich eine höhere zusätzliche Maskierung $\Delta L_T(z_k, m)$. Wie in Abbildung 6.14 sichtbar ist, haben die Auflösung Δz und der Parameter a auf die psychoakustische Optimierung des nichtlinearen Modells großen Einfluß. Die besten Ergebnisse wurden mit $\Delta z = 0{,}43$ und $a = 0{,}3$ erreicht. Das belegen auch psychoakustische Daten aus [110].

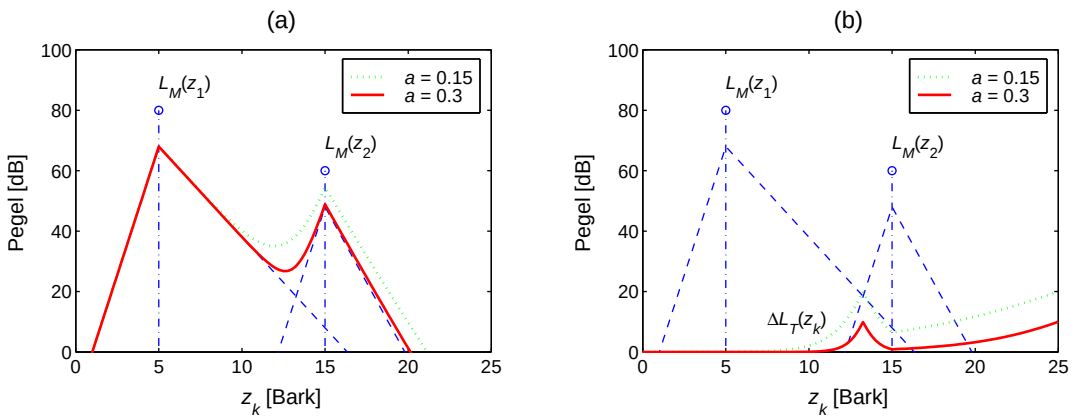


Abbildung 6.14 Nichtlineare Superposition zweier lokaler Maskierer mit den Maskierschwellen $L_{T,1}(z_k, m)$ und $L_{T,2}(z_k, m)$. **(a)** Die resultierende globale Mithörschwelle $L_T(z_k, m)$ ist für unterschiedliche a darstellt. **(b)** Die zusätzliche Maskierung wird mit $\Delta L_T(z_k, m)$ bezeichnet.

6.3.5 Inverse Filterung

Nach Superposition der lokalen Maskierschwellen erfolgt zur Berechnung der globalen Maskierschwelle $L_T(z_k, m)$ die inverse Filterung mit $\Xi^{-1}(z_k)$. Vor- und Nachfilterung bewirken die gehörrichtige Gewichtung der einzelnen lokalen Maskierer. Durch die inverse Filterung wird der Einfluß des Vorfilters aus Absatz 6.3.2 kompensiert und es ergibt sich die Maskierschwelle $L_T(z_k, m)$, die auf ein gegebenes Eingangssignal wirkt. Für die resultierende globale Maskierschwelle folgt:

$$L_T(z_k, m) = I_T(z_k, m) \cdot \Xi^{-1}(z_k). \tag{6.28}$$

Nach inverser Filterung liegt die globale Maskierschwelle $L_T(z_k, m)$ auf der Tonheitsskala z_k vor. Durch Abbildung auf die übliche diskrete Frequenzskala mit den Frequenzstützstellen n erhält man die globale „diskrete“ Maskierschwelle $L_T(n, m)$ (in dB) mit der spektralen Mithörschwelle $R_T(n, m)$ zu:

$$R_T(n, m) = 10^{\frac{L_T(n,m)}{10}}. \tag{6.29}$$

Die berechnete Mithörschwelle $R_T(n, m)$ stellt keine Leistungsdichte dar, ist aber durch psychoakustische Verdeckungseffekte eng mit dem Leistungsdichtespektrum $R_{ss}(n, m)$ des Sprachsignals verknüpft, vgl. Abschnitt 3.3.

6.4 Schätzung der Leistungsdichte des Nutzsignals

In [13], [14], [71] und [72] sind zur Schätzung der psychoakustischen Maskierschwellen die Kurzzeit-Leistungsdichtespektren des gestörten Sprachsignals $x(k)$ verwendet worden. Dadurch ergaben sich besonders bei kleinem Signal-Störabstand häufig zu optimistische Mithörschwellen. Um die tatsächlichen vom Sprachsignal $s(k)$ hervorgerufenen spektralen Mithörschwellen $R_T(n, m)$ zu bestimmen, wird deshalb von der Schätzung der Leistungsdichte $\tilde{R}_{ss}(n, m)$ des Sprachsignals $s(k)$ ausgegangen.

Für die Bestimmung der spektralen psychoakustischen Mithörschwelle $R_T(n, m)$ gemäß Abschnitt 6.3 ist die Kenntnis der spektralen Leistungsdichte $R_{ss}(n, m)$ des störungsfreien Sprachsignals $s(k)$ erforderlich, die nur aus dem vorliegenden Signalgemisch $x(k)$ geschätzt werden kann. Bereits in Abschnitt 3.1.1 wurde auf die Merkmale und Analyse des Sprachsignals eingegangen. Die dort beschriebenen Methoden sind der Ausgangspunkt und liefern

Ansätze für die Schätzung der Sprachsignalleistungsdichte $\hat{R}_{ss}(n, m)$. Wegen der Instationarität von Sprache ist eine Beschreibung und Analyse des Sprachsignals nicht mit Hilfe klassischer Verfahren möglich. Dennoch gelingt es, besondere Merkmale der Sprache zu extrahieren, wenn die Analyse des Sprachsignals nur für einen kurzen Zeitabschnitt vorgenommen wird (Kurzzeitanalyse).

6.4.1 Schätzung mit spektraler Subtraktion

Die Schätzung der ungestörten Nutzsignalleistungsdichte in (6.30) führt auf die Anwendung der spektralen Subtraktion zurück. Mit (2.100) ergibt sich für die Schätzung der Nutzsignalleistungsdichte bei einem orthogonalen, statistisch unabhängigen Störer $n(k)$:

$$\begin{aligned}\hat{R}_{ss}(n,m) &= R_{xx}(n,m) - \hat{R}_{nn}(n,m) \\ &= R_{xx}(n,m)\cdot\left[1-\frac{\hat{R}_{nn}(n,m)}{R_{xx}(n,m)}\right] \\ &= R_{xx}(n,m)\cdot\left|G(n,m)\right|^2 .\end{aligned} \tag{6.30}$$

Die Kurzzeit-Leistungsdichte des Störsignals ist üblicherweise nicht frei zugänglich. Deshalb wird hier wieder die Schätzung $\hat{R}_{nn}(n, m)$ verwendet. Durch den Einsatz einer einfachen spektralen Subtraktion kommt es bei Fehlschätzung der Leistungsdichte des Störsignals besonders bei kleinem SNR zu sporadisch auftretenden Spektrallinien, die als „musical tones“ bereits beschrieben wurden. Diese Reststörungen verändern den tonalen Charakter des Nutzsignals. In Abschnitt 6.6 wird zudem untersucht, welche Auswirkungen diese Effekte auf die Bestimmung der globalen psychoakustischen Mithörschwelle haben. Dazu wird das Gesamtsystem in Abhängigkeit von verschiedenen Verfahren und Algorithmen für die Nutzsignal- und Störsignalschätzung bewertet und bezüglich der Leistungsfähigkeit der Geräuschreduktion und der Verbesserung der Sprachverständlichkeit optimiert.

6.4.2 Schätzung durch lineare Prädiktion

Die lineare Prädiktion ermöglicht eine äußerst kompakte Kurzzeitrepräsentation des abgetasteten Sprachsignals mit relativ geringem Aufwand. Der allgemeine Prädiktor liefert aus vorhergehenden Abtastwerten eine Schätzung für den folgenden Abtastwert (Extrapolation, Vorwärtsprädiktion) bzw. aus noch bekannten Signalwerten eine Schätzung für einen vergangenen Wert (Interpolation, Rückwärtsprädiktion). Die Identifikation des Vokaltraktfilters anhand der linearen Prädiktion ermöglicht die Extraktion geringdimensionaler Sprachmerkmale, die im Bereich der Mustererkennung, Sprachsynthese und Codierung erfolgreich einge-

setzt werden. Ausgangspunkt für die Analyse des Sprachsignals anhand der linearen Prädiktion ist das in Abschnitt 3.1.1.1 vorgestellte zeitdiskrete Modell der Spracherzeugung. In der Realität ist von gemischter Anregung auszugehen, siehe Abbildung 3.2. Zur Vereinfachung wurde anstelle eines Summationsglieds ein Schalter eingeführt, der zwischen stimmhafter und stimmloser Anregung umschaltet. Diese Vereinfachung hat sich für die parametrische Beschreibung der Sprachübertragung bewährt.

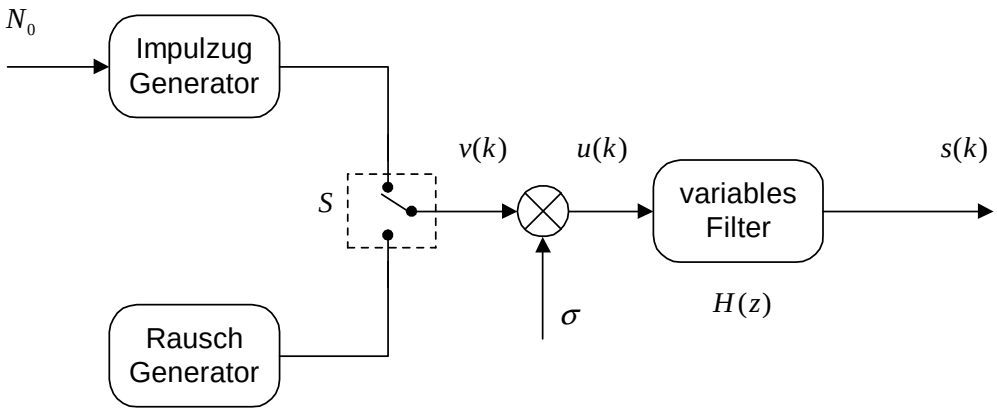


Abbildung 6.15 Zeitdiskretes Modell der Spracherzeugung. Ausgegangen wurde vom Fant'schen Source-Filter Modell lt. Abbildung 3.2, wobei das Summierglied hier durch einen Schalter S ersetzt wurde, der zwischen stimmloser und stimmhafter Anregung verzweigt.

Das an sich zeitvariable Vokaltraktfilter $h(k)$ wird zunächst als zeitinvariant behandelt. Durch z-Transformation ergibt sich die Übertragungsfunktion des Vokaltrakts $H(z)$. Der Vokaltrakt wird mit dem Signal

$$u(k) = \sigma v(k) \tag{6.31}$$

angeregt, wobei der Verstärkungsfaktor σ die Amplitude des Anregungssignals $u(k)$ bestimmt. Zur Synthese *stimmhafter* Abschnitte wird eine δ-Impulsfolge (Dirac)

$$v(k) = \sum_{i=-\infty}^{\infty} \delta(k - iN_0) \tag{6.32}$$

mit der Periode N_0 verwendet. Zur Synthese *stimmloser* Abschnitte kommt ein weißes Rauschsignal $v(k)$ mit der Varianz $\sigma_v^2 = 1$ zum Einsatz. Im allgemeinen Fall ist der Zusammenhang zwischen dem Anregungssignal $u(k)$ und dem Ausgangssignal $s(k)$ im Zeitbereich durch die allgemeine Differenzengleichung eines digitalen Filters I-ten Grades gegeben:

$$s(k) = \sum_{i=0}^{I} a_i u(k-i) - \sum_{i=0}^{I} b_i s(k-i). \tag{6.33}$$

Nach z-Transformation erhält man die allgemeine zeitinvariante Übertragungsfunktion des Vokaltrakts als:

$$H(z) = \frac{S(z)}{U(z)} = H_0 \cdot \frac{1 + \sum_{i=1}^{I} a_i z^{-i}}{1 - \sum_{i=1}^{I} b_i z^{-i}}. \tag{6.34}$$

Der allgemeine Fall des *Pol-Nullstellen-Modells (Auto Regressive Moving-Average, ARMA-Modell)* wird durch die Gleichungen (6.33) und (6.34) erfaßt. Für $b_i = 0$ mit $i = 1, 2, \ldots, M$ wird das Filter in (6.34) nicht rekursiv. Diesen Fall bezeichnet man *Nullstellen-Modell (Moving-Average, MA-Modell)*. Die Übertragungsfunktion wird nur durch ihre Nullstellen bestimmt. Für $a_i = 0$, wobei $i = 1, 2, \ldots, I$, entsteht mit

$$s(k) = a_0 u(k-I) - \sum_{i=1}^{I} b_i s(k-i) \tag{6.35}$$

ein rekursives Filter im Zeitbereich bzw. mit

$$H(z) = \frac{a_0 z^{-I}}{1 - \sum_{i=1}^{I} b_i z^{-i}} \tag{6.36}$$

ergibt sich ein Filter im Frequenzbereich, das abgesehen von einer I-fachen Nullstelle bei $z = 0$ nur Pole besitzt.

Damit erhält man die allgemeine Form eines Allpolfilters und man spricht von einem *Autoregressiven Prozeß (AR-Modell)*. Dieser Prozeß entspricht dem Modell der Spracherzeugung aus Abschnitt 3.1.1.

Ausgehend vom Fant'schen Source-Filter Modell kann der Sprachprozeß durch Hintereinanderschaltung mehrerer linearer zeitvarianter Systeme modelliert werden. Dabei wird das Sprachsignal im wesentlichen durch die sich zeitlich rasch ändernde Anregung $U(z)$ und die

relativ langsame Signalformung im Vokaltrakt $H(z)$ bestimmt. Zur Analyse des Sprachsignals $s(k)$ sind die Koeffizienten b_i aus Gleichung (6.35) zu schätzen, d.h. der Vokaltrakt $H(z)$ wird identifiziert. Die Bestimmung der Prädiktionskoeffizienten entspricht der Identifikation der Vokaltraktübertragungsfunktion $H(z)$. Das Pol-Nullstellenfilter sei kausal und stabil. In der Z-Ebene liegen damit die Polstellen innerhalb des Einheitskreises, während die Nullstellen auch außerhalb des Einheitskreises liegen können. Jedes derartige System $H(z)$ läßt sich in ein minimalphasiges Filter und ein Allpaßsystem mit

$$H(z) = H_{\min}(z) H_{Ap}(z) \tag{6.37}$$

aufspalten. In Abbildung 6.16 wird dies für ein System mit zwei Polstellen-Nullstellen-Paaren verdeutlicht.

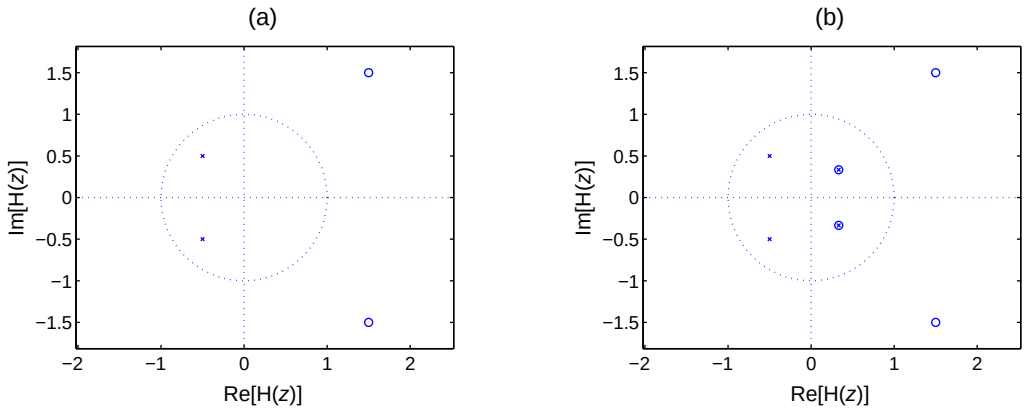


Abbildung 6.16 Pol-Nullstellen Diagramm der Vokaltraktübertragungsfunktion. **(a)** Ursprüngliches Pol-Nullstellen-Filter. **(b)** Aufspaltung in minimalphasiges Filter und Allpaß

Für die psychoakustische Sprachsignalanalyse genügt es, lediglich den minimalphasigen Anteil zu untersuchen, da das Ohr gegenüber der durch den Allpaß hervorgerufenen Veränderung der Phase weitgehend unempfindlich ist, siehe Abschnitt 3.2.2.3. Daraus ergeben sich zwei wichtige Konsequenzen:

- Die durch den Sprechtakt hervorgerufene Filterung kann prinzipiell durch inverse Filterung des Sprachsignals rückgängig gemacht werden, so daß sich das Anregungssignal des Sprechtraktes zurückgewinnen läßt.
- Da Pole und Nullstellen des minimalphasigen Filters innerhalb des Einheitskreises liegen, existiert ein stabiles inverses Filter mit

$$H_{\min}^{-1}(z) = \frac{1}{H_{\min}(z)}. \tag{6.38}$$

- Jedes minimalphasiges Pol-Nullstellen-Filter kann exakt durch ein Allpol-Filter mit unendlich hohem Grad dargestellt werden und durch ein Filter M-ten Grades approximiert werden. Dadurch ist die Verwendung des Allpol-Filters für die Modellierung des Sprachsignals geeignet.

Die Koeffizienten des Allpol-Filters $H_{min}(z)$ lassen sich mit der Technik der linearen Prädiktion bestimmen. Die lineare Prädiktion impliziert, daß aufeinanderfolgende Abtastwerte $s(k)$ aufgrund der Filterung eine statistische Abhängigkeit aufweisen. Der Abtastwert $s(k)$ wird bei gegebenen Koeffizienten b_i bis auf die sogenannte Innovation $u(k)$ durch die vorhergehenden Abtastwerte $s(k-1), s(k-2), \ldots, s(k-I)$ bestimmt.

Ausgangspunkt für die Bestimmung der Prädiktorkoeffizienten bilden die Gleichungen (6.35) und (6.36). Da die Modellkoeffizienten b_i nicht bekannt sind, wird folgende Schätzung für $s(k)$ angesetzt:

$$\hat{s}(k) = \sum_{i=1}^{I} b_i s(k-i). \tag{6.39}$$

Diese Form der Schätzung wird als *lineare Prädiktion* mit dem Prädiktionsfehlersignal

$$d(k) = s(k) - \hat{s}(k) \tag{6.40}$$

bezeichnet. Die Prädiktorkoeffizienten b_i sollen im Sinne des minimalen mittleren quadratischen Fehlers optimiert werden:

$$E\left\{d^2(k)\right\} = E\left\{\left[s(k) - \sum_{i=1}^{I} b_i s(k-i)\right]^2\right\} \rightarrow \min. \tag{6.41}$$

Zur Vereinfachung sollen vorerst folgende Annahmen gelten: Die unbekannten Prädiktorkoeffizienten b_i und die Vokaltraktübertragungsfunktion $H(z)$ seien zeitinvariant. Das Anregungssignal $u(k)$ sei reell, stationär, unkorreliert und mittelwertfrei (weißes Rauschen). Die optimalen Prädiktorkoeffizienten b_i mit $i = 0, 1, \ldots, I$ lassen sich aus Beziehung (6.41)

durch Bildung der Ableitung nach einem der Koeffizienten b_i bestimmen. Die partielle Ableitung des mittleren quadratischen Fehlers nach dem festen Koeffizienten b_i liefert mit (6.39) und (6.40):

$$\begin{aligned}\frac{\partial E\left\{d^2(k)\right\}}{\partial b_i} &= \frac{\partial E\left\{\left[s(k)-\sum_{i=1}^{I} b_i s(k-i)\right]^2\right\}}{\partial b_i} \\ &= -2E\left\{\left(s(k)-\sum_{i=1}^{I} b_i s(k-i)\right)s(k-j)\right\} \quad \text{mit } j=1,2,\ldots,I \\ &= -2E\left\{d(k)s(k-j)\right\}=0.\end{aligned} \tag{6.42}$$

Gleichung (6.42) gibt das bereits in Abschnitt 2.5 erläuterte Orthogonalitätsprinzip wieder. Danach ist bei optimalen Prädiktionskoeffizienten b_i das Prädiktionsfehlersignal $d(k)$ gemäß Gleichung (6.40) orthogonal zu dem Eingangssignal $s(k)$. Mit Verwendung der Autokorrelationsfunktion $r_{ss}(i)$ nach Definition (2.30) folgt die sogenannte *Yule-Walker-Prädiktionsgleichung:*

$$-r_{ss}(j)+\sum_{i=1}^{I} b_i r_{ss}(j-i)=0. \tag{6.43}$$

Für $j = 1, 2, \ldots, I$ ergeben sich die sogenannten Normalengleichungen

$$\begin{pmatrix} r_{ss}(1) \\ r_{ss}(2) \\ \\ r_{ss}(I) \end{pmatrix} = \begin{pmatrix} r_{ss}(0) & r_{ss}(-1) & \cdots & r_{ss}(1-I) \\ r_{ss}(1) & r_{ss}(0) & \cdots & r_{ss}(2-I) \\ & & \cdots & \\ r_{ss}(I-1) & r_{ss}(I-2) & \cdots & r_{ss}(0) \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \\ \\ b_I \end{pmatrix} \tag{6.44}$$

bzw. in Kurzform:

$$\mathbf{q} = \mathbf{r_{ss}}\mathbf{b}. \tag{6.45}$$

Als Lösung der Normalengleichungen in (6.44) bzw. (6.45) erhält man den optimalen Koeffizientenvektor $\mathbf{c_{LP}} = \mathbf{b}$ zu

$$\mathbf{c_{LP}} = \mathbf{r_{ss}^{-1}q} = \left[c_{LP}(1), c_{LP}(2), ..., c_{LP}(I)\right]^T . \tag{6.46}$$

Der Koeffizientenvektor $\mathbf{c_{LP}}$ stellt die optimale Schätzung der Prädiktorkoeffizienten b_i dar. Das Vokaltraktsystem $H(z)$ lt. Gleichung (6.35) wurde identifiziert. In Abbildung 6.17 wird deutlich, daß bereits I Prädiktionskoeffizienten zu einer guten Approximation des tatsächlichen Spektralverlaufs der Sprache führen. Dazu wurde ein Sprachsignal im Fahrzeug aufgezeichnet, der LPC-Analyse mit I Koeffizienten unterzogen und anschließend synthetisiert. Das geschätzte Leistungsdichtespektrum des Sprachsignals wird folgendermaßen bestimmt, wobei bei Anregung mit weißem Rauschen $R_{uu}(\Omega) = 1$ gilt:

$$\begin{aligned}\hat{R}_{ss}(\Omega) &= R_{uu}(\Omega) \cdot \left|H(\Omega)\right|^2 \\ &= \left| \frac{a_0 e^{-jI\Omega}}{\sum_{i=1}^{I} b_{I-i} e^{-j\Omega i}} \right|^2 , \quad \text{mit } \Omega = 2\pi \frac{f}{f_a}.\end{aligned} \tag{6.47}$$

In [44] wird für die Dimensionierung der Filterordnung I

$$\mathrm{I} = \begin{cases} \left.\dfrac{f_a}{\mathrm{kHz}}\right|_{\text{stimmhaft}} \\ \left.\dfrac{f_a}{\mathrm{kHz}} + 4\right|_{\text{stimmlos}} \end{cases} \tag{6.48}$$

vorgeschlagen. Damit steht pro kHz sogenannter Nyquistfrequenz $f_a/2$ ein Polpaar der Übertragungsfunktion des Vokaltrakts $H(\Omega)$ zur Verfügung. Das entspricht der Vorstellung, daß das Spektrum des Sprachsignals etwa einen Formanten pro kHz besitzt. Für die Modellierung der stimmlosen Sprachanteile werden vier weitere Pole genutzt, um die glottale Anregung voll zu erfassen. Die lineare Prädiktion ermöglicht eine äußerst kompakte Kurzzeitrepräsentation des abgetasteten Sprachsignals mit relativ geringem Aufwand.

Für die Beachtung des instationären Verhaltens des Vokaltraktes kommt die Kurzzeitanalyse durch Einführung des Fensterindex $m = 1, 2, \ldots, M$ und des Frequenzindex $n = 1, 2, \ldots, N$ zum Einsatz. Das geschätzte Leistungsdichtespektrum des Sprachsignals wird dann folgendermaßen bestimmt:

$$\hat{R}_{ss}(n,m) = R_{uu}(n,m)\cdot\left|H(n,m)\right|^2 = \left|\frac{a_0(m)\cdot e^{-j\frac{2\pi}{N}I}}{\sum_{i=1}^{I} b_{I-i}(m)\cdot e^{-j\frac{2\pi}{N}i}}\right|^2 . \tag{6.49}$$

Dabei ergibt sich bei Anregung mit weißem Rauschen: $R_{uu}(n,m) = 1$. Geht man davon aus, daß die Störanteile $n(k)$ im Signalgemisch $x(k)$ nicht als AR-Prozeß geringer Dimension modelliert werden können, so läßt sich Ausdruck (6.49) als Schätzung der Leistungsdichte des Sprachsignals in gestörter Umgebung interpretieren.

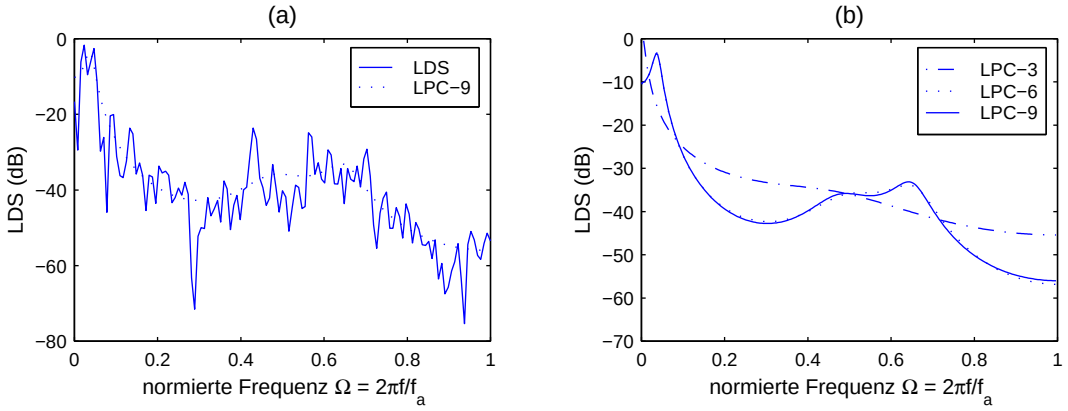


Abbildung 6.17 (a) Leistungsdichtespektrum eines im BMW 528i aufgezeichneten Sprachsignalsausschnitts mit ca. 15 ms Dauer. **(b)** LPC-Analyse (Korrelationsmethode, vgl. [183]) und anschließende Rücksynthese des Ursprungssignals. Ab LPC-Filterordnung $I = 6$ sind kaum noch Verbesserungen der Synthese des Originalsignals wahrzunehmen. Die Formantstruktur des ursprünglichen Sprachsignals wird bereits gut reproduziert.

6.4.3 Schätzung mit Tiefpaß-Lifterung

Eine weitere Möglichkeit der Sprachsignalschätzung geht erneut auf das Fant'sche Source-Filter Modell aus Abschnitt 3.1.1.1 zurück. Hierbei wird versucht, mittels homomorpher Entfaltung Anregung $U(z)$ und Signalformung $H(z)$ zu trennen. Nach [125] läßt sich der Verlauf des komplexen Kepstrums $C_s(\kappa)$ des Sprachsignals $s(k)$ als

$$C_s(\kappa) = \mathbf{F^{-1}}\left\{\log S(e^{j\Omega})\right\} \text{ mit } S(e^{j\Omega}) = \mathbf{F}\left\{s(k)\right\} \tag{6.50}$$

darstellen. Die Variable κ wird als Quefrenz bezeichnet. Die Logarithmierung überführt die Multiplikation aus Gleichung (3.1) in eine additive Verknüpfung. Ist das Signal durch Faltung

entstanden, lassen sich die Anteile als Summanden im logarithmierten Spektrum wiederfinden. Nach inverser Transformation zurück in den Frequenzbereich bleibt wegen der Linearität der Summation die additive Überlagerung erhalten. Bereits in [130] wurde auf die Bildung der Mel-Kepstral-Koeffizienten mittels einer Filterbank eingegangen. Dabei wird das Sprachsignal $s(k)$ in eine I-dimensionale, psychoakustische *Bark*-Filterbank eingespeist. Die Mel-Energie aus den einzelnen Filtern wird dann logarithmiert und der inversen Cosinustransformation unterzogen. So entsteht ein sehr kompakter I-dimensionaler Merkmalsvektor (*Mel-Cepstral-Vektor*), der sich in der Sprachverarbeitung und Codierung bewährt hat.

Durch die Begrenzung der Dimension, üblicherweise auf $I = 10 \ldots 22$, werden nur die unteren Kepstralkoeffizienten ausgewertet[1]. Die *Kepstral-Koeffizienten* $c_{cc}(i)$ lassen sich nach [59] auch durch Rekursion aus den LP-Koeffizienten c_{LP} wie folgt bestimmen:

$$\begin{aligned} c_{CC}(1) &= -c_{LP}(1) \\ c_{CC}(i) &= -c_{LP}(i) - \sum_{k=1}^{i-1}\left(1-\frac{k}{i}\right)\cdot c_{LP}(k)\cdot c_{CC}(i-k). \end{aligned} \tag{6.51}$$

Sei $s(k)$ das störungsfreie Nutzsignal mit der Kurzzeitleistungsdichte $\tilde{R}_{ss}(n, m)$. Die diskrete Cosinus-Tranformation der gehörrichtigen logarithmierten Schallintensitäten aus (6.17) ergibt die *Mel-Frequenz-Kepstral-Koeffizienten (MFCC)*:

$$\mathbf{c_q(q}, m\mathbf{)} = 2\cdot\sum_{k=0}^{K-1}\log R_{ss}\left(\frac{2\pi k\mathbf{q}}{K}, m\right)\cos\frac{2\pi k\mathbf{q}}{K},\ \mathbf{q} = [0, 1, \ldots, K-1]^T. \tag{6.52}$$

Verringert man die Dimension des MFCC-Vektors von K auf die ersten K_{LT} Komponenten, so erhält man einen sehr kompakten Merkmalsvektor $\boldsymbol{c_{LT}}$. Wegen der geringen Dimension und der Kompaktheit hat sich der MFCC-Vektor als Sprachmerkmal für den Einsatz in Spracherkennungssystemen durchgesetzt. Der reduzierte Merkmalsvektor ergibt sich aus:

$$\mathbf{c_{LT}(q}, m\mathbf{)} = \sum_{q=0}^{K_{LT}} c_q(q, m). \tag{6.53}$$

[1] Dieses Verfahren wird als Tiefpaß-Lifterung bezeichnet. Als Hochpaß-Lifterung wird dementsprechend die Verwendung des oberen Teils des Mel-Kepstral-Vektors benannt. Dies kann mit der Analyse der Sprachanregung assoziiert werden. Man beachte hier die besondere Nomenklatur: Das Kepstrum kann genauso als die Fouriertransformation eines Spektrums verstanden werden. Die Ähnlichkeit zu einer Spektraltransformation hat man durch die Bezeichnung deutlich gemacht. Dabei erhält man durch das Vertauschen der Buchstaben im Wort „Spektrum“ das „Kepstrum“. Genauso ersetzt man „Frequenz“ durch „Quefrenz“ und vertauscht „Filterung“ mit „Lifterung“.

Die Reduzierung der Dimension des MFCC mit anschließender inverser diskreter Cosinus-Transformation (*IDCT*) bezeichnet man als *Tiefpaß-Lifterung*. Durch die Tiefpaß-Lifterung werden die Anregungsanteile der Sprache von den Vokaltraktanteilen getrennt. Führt man die Tiefpaßlifterung an gestörten Sprachsignalen $x(k)$ durch, so erhält man die Schätzung der Leistungsdichte $\hat{R}_{ss}(n, m)$ des mittelwertfreien Nutzsignals $s(k)$ mit:

$$\hat{R}_{ss}(n,m) = \frac{1}{K}\sum_{q=0}^{K-1} 10^{c_{LT}(q,m)} \cos(nq). \tag{6.54}$$

Mit homomorpher Entfaltung durch Kepstral-Analyse und Tiefpaß-Lifterung läßt sich damit das Leistungsdichtespektrum des ungestörten Sprachsignals schätzen.

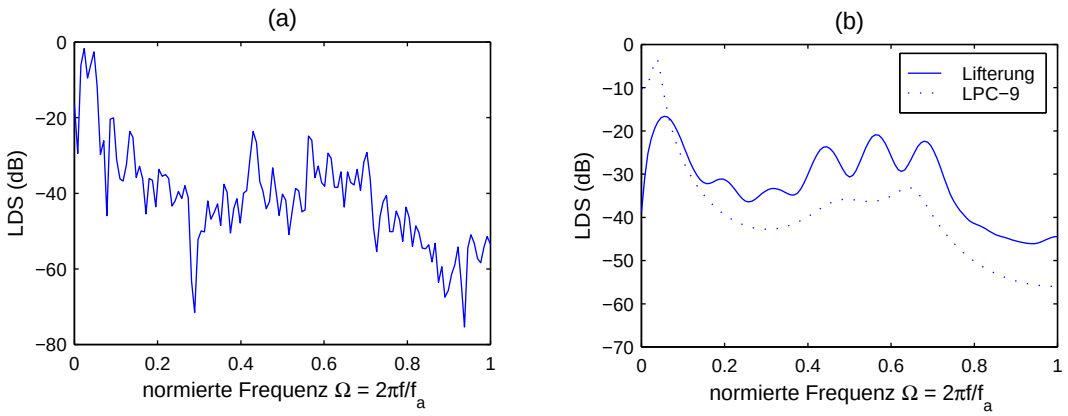


Abbildung 6.18 Vergleich zwischen Leistungsdichtespektrum eines im BMW 528i augezeichneten Sprachsignals **(a)** und LDS des Signals nach kepstraler Tiefpaß-Lifterung mit $I = 16$ **(b)**, sowie nach LPC-Analyse mit neun Koeffizienten. Deutlich zeigt sich die Rekonstruktion der Formantstruktur des Sprach-Signals durch die Lifterung.

6.4.4 Diskussion und Vergleich der Verfahren

Analog Gleichung (6.12) wurde der absolute Fehler $\left|P_e(SNR)\right|$ zwischen dem Kurzzeit-Leistungsdichtespektrum $\tilde{R}_{ss}(n, m)$ des Sprachsignals und seiner Schätzung $\hat{R}_{ss}(n, m)$ mit

$$\left|P_e(SNR)\right| = \frac{1}{M \cdot N}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\left|\frac{\hat{R}_{ss}(n,m) - \tilde{R}_{ss}(n,m)}{\tilde{R}_{xx}(n,m)}\right|_{SNR} \tag{6.55}$$

berechnet und für verschiedene Signal-Störabstände ausgewertet. Abbildung 6.19 zeigt den Vergleich der Verläufe des absoluten Fehlers $\left|P_e(SNR)\right|$ für unterschiedliche Störszenarien.

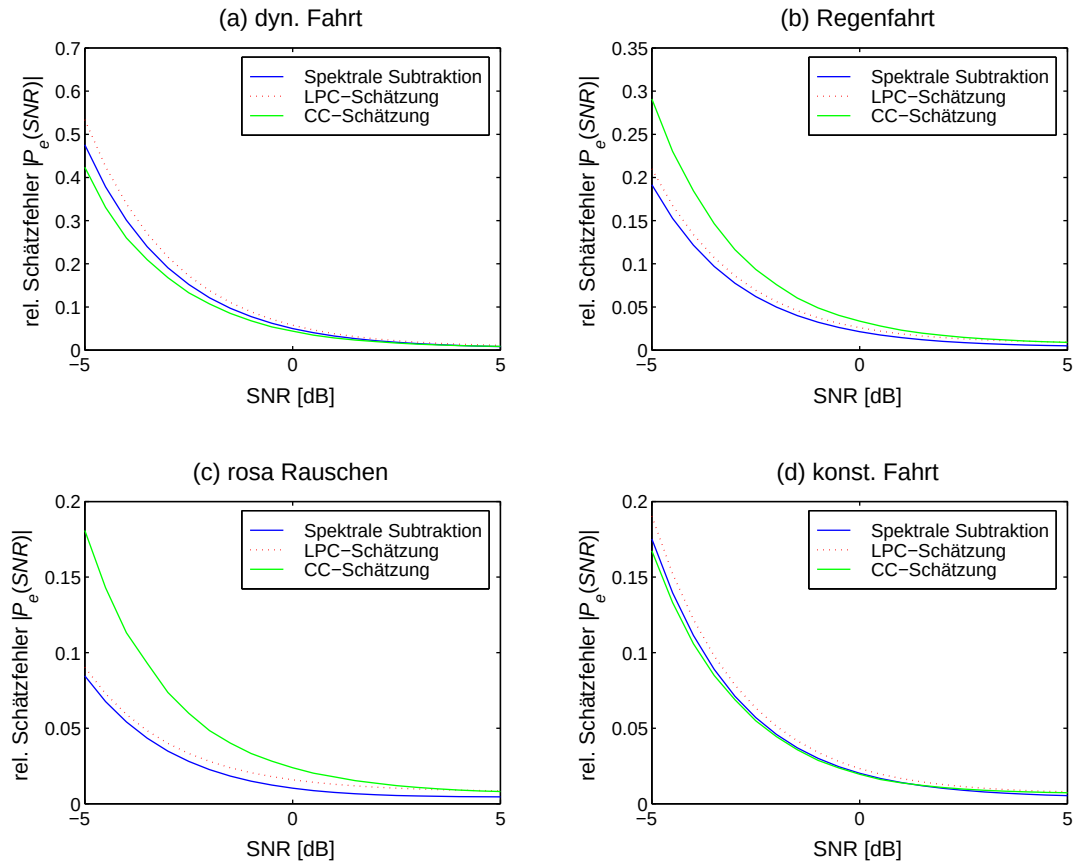


Abbildung 6.19 Vergleich der Verfahren zur Schätzung der Leistungsdichte $\hat{R}_{ss}(n, m)$ des Sprachsignals bei unterschiedlichen Signal-Störverhältnissen und verschiedenen Geräuschszenarien. Aufnahme der Geräusche und des Sprachsignals im BMW 528i touring.

Dabei fällt zunächst die Überlegenheit der einfachen spektralen Subtraktion gemäß Abschnitt 6.4.1 auf. Sie ergibt für alle Störgeräusche den kleinsten absoluten Schätzfehler. Dagegen wird der Unterschied zwischen der LPC- und CC-Schätzung erst bei negativem SNR deutlich.

Dennoch ist anzunehmen, daß mit dem LPC- und CC-Verfahren die Rekonstruktion der Formantstruktur des Sprachspektrums besser gelingt als mit der einfachen spektralen Subtraktion. Sowohl die lineare Prädiktion wie auch die homomorphe Entfaltung mit dem Kepstralverfahren lehnt sich an die in Abschnitt 3.1.1.3 behandelten Modelle der Spracherzeugung im menschlichen Vokaltrakt an.

Weiterhin ist zu bedenken, daß sich bei Simultanverdeckung die psychoakustischen Verdekkungsschwellen besonders in Nähe energiereicher Formanten[1] ausbilden. Deshalb ist es auch wahrscheinlich, daß sich diese modellorientierten Verfahren besonders gut für die Bestimmung der Mithörschwellen eignen, obwohl sie keine besseren Schätzungen der Nutzsignallei-

[1] Die Resonanzfrequenzen des Vokaltrakts werden als Formanten bezeichnet.

stungsdichte liefern als die spektrale Subtraktion. Um diese Vermutung zu bestätigen, wurde der Vergleich der vorgestellten Verfahren zur Schätzung der Nutzsignalleistungsdichte geändert.

Zunächst wurden die Formantstruktur und die Leistungsdichte eines ungestörten stationären Sprachsignals $s(k)$ vorgegeben. Das Signal $s(k)$ enthält drei sinusförmige Töne bei 215, 646 und 1292 Hz mit gleicher Leistung. Das Spektrogramm dieses Signalgemisches $s(k)$ ist in Abbildung 6.20 dargestellt. Hierbei wurde die Spektrale Subtraktion mit dem CA-Schätzverfahren, vgl. Abschnitt 6.2.3, durchgeführt. LPC- wie auch CC-Schätzung wurden mit $I = 16$ durchgeführt. Schon am ungestörten Testsignal zeigt sich, daß sowohl das LPC- wie auch das CC-Verfahren die Spitzen im Leistungsdichtespektrum $R_{ss}(n)$ besser rekonstruieren.

Dagegen ergibt sich mit dem spektralen Subtraktionsverfahren eine bessere Schätzung des Verlaufs der Leistungsdichte des Testsignals $s(k)$. Daher rührt auch der gemessene kleinere Fehler über den gesamten Vergleich zwischen dem Leistungsdichtespektrum $R_{ss}(n)$ und der Schätzung $\hat{R}_{ss}(n)$. Dagegen schätzt das CC-Verfahren den Verlauf von $R_{ss}(n)$ im Vergleich zum LPC-Verfahren zu ungenau. Für die Bestimmung der psychoakustischen Mithörschwellen ist vor allem die exakte Rekonstruktion der Formanten wichtig. Der Verlauf zwischen den Formanten wird oftmals durch Automaskierung verdeckt.

Um diese Rekonstruktion auch für gestörte Signale zu untersuchen, erfolgte anschließend der Vergleich der drei Verfahren für geräuschbehaftete Umgebungen, siehe Abbildung 6.21. Bei allen Störgeräuschen wird deutlich, daß es der LPC-Schätzung besser als der CC-Schätzung oder der spektralen Subtraktion gelingt, die Formanten zu rekonstruieren. Dabei ergibt sich in Spitze ein Schätzfehler von bis zu 6 dB. Inwieweit sich die verschiedenen Verfahren auf die Bestimmung der Maskierschwellen auswirken, wird gesamthaft in Abschnitt 6.6 untersucht.

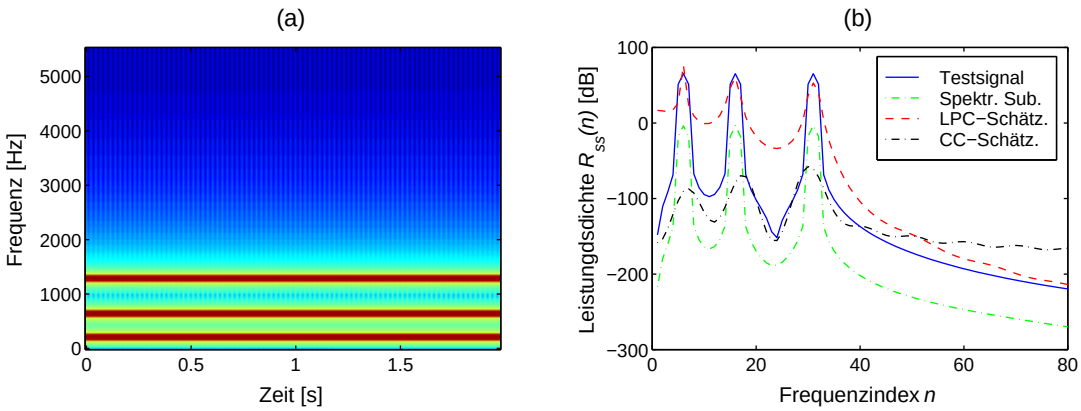


Abbildung 6.20 Formantstruktur und Rekonstruktion eines exemplarischen Testsignals aus drei gemischten Sinustönen bei 215 ($n = 5$), 646 ($n = 15$) und 1292 Hz ($n = 30$). Abtastfrequenz $f_a = 11025$ Hz. **(a)** Spektrogramm des Testsignals $s(k)$. **(b)** Leistungsdichte $R_{ss}(n)$ des Testsignals und Schätzungen der Leistungsdichte mit drei verschiedenen Verfahren.

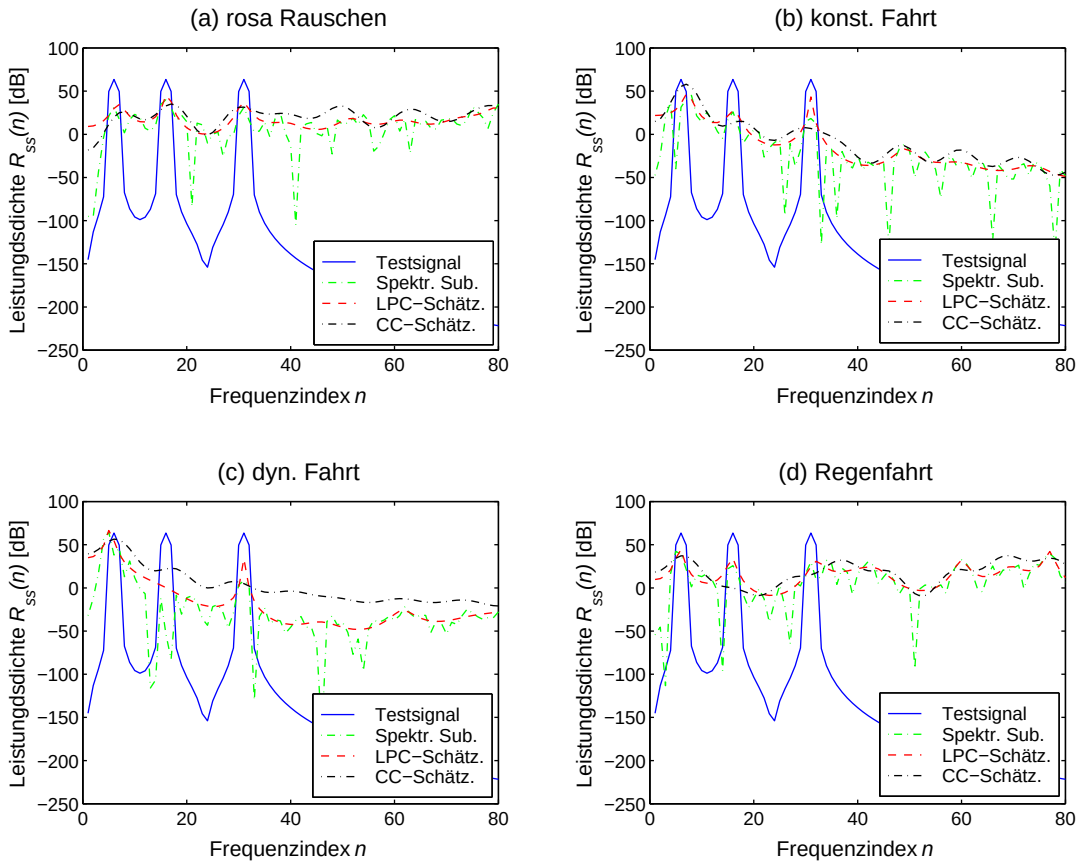


Abbildung 6.21 Rekonstruktion der Formantstruktur in geräuschbehafteten Testsignalen. Aufnahme der Geräusche im BMW 528i touring. $SNR = 0$ dB. **(a)** Rosa Rauschen, **(b)** Fahrt mit konstanter Geschwindigkeit, **(c)** beschleunigte Fahrt und **(d)** Fahrt im Regen.

6.5 Bestimmung der optimalen Gewichtsfunktion

Gemäß Abbildung 6.1 wird das adaptive Filter $G(n, m)$ gesucht, das mit

$$\hat{S}(n,m) = X(n,m) \cdot G(n,m) \tag{6.56}$$

eine optimale Schätzung von $S(n, m)$ vornimmt. Die Übertragungsfunktion $G(n, m)$ wird in Abhängigkeit von der globalen Mithörschwelle $R_T(n, m)$ bestimmt. Das Ziel ist, das Signal-Stör-Verhältnis zu verbessern und dabei die Verzerrungen des Nutzsignals und der Restgeräusche unter die Mithörschwelle zu bringen, so daß sie verdeckt werden. Dabei ist es besonders wichtig, die resultierende Sprachverständlichkeit zu erhöhen und den subjektiven Höreindruck des entstörten Signals zu verbessern.

6.5.1 Parametrische spektrale Subtraktion

Mit Parametrisierung von (6.56) ergibt sich für das geschätzte Spektrum des Nutzsignals $s(k)$:

$$\hat{S}(n,m) = S(n,m) \cdot G_W(n,m) + N(n,m) \cdot \left[b - a \cdot G_W(n,m) \right] \tag{6.57}$$

mit der Übertragungsfunktion des bekannten Wiener-Filters:

$$G_W(n,m) = \left(\frac{R_{ss}(n,m)}{R_{xx}(n,m)} \right)^{\gamma} = \left(\frac{\xi(n,m)}{1+\xi(n,m)} \right)^{\gamma} . \tag{6.58}$$

Diese allgemeine Form der spektralen Subtraktion erlaubt durch Variation der Parameter a, b, γ einen Kompromiß zwischen Geräuschreduktion, Reststörung und Verzerrung des Nutzsignals. Die eingeführten Parameter haben folgende Bedeutung und Wirkung:

Der sogenannte *Oversubtraction* Faktor *a*, mit

$$0 < a \leq 1 , \tag{6.59}$$

bewirkt die Erhöhung der Dämpfung des Störgeräusches hat aber gleichzeitig eine stärkere Verzerrung des Nutzsignals zur Folge. Der *Spectral Floor b*, mit

$$0 < b < 1 , \tag{6.60}$$

ermöglicht das teilweise Verbleiben des Hintergrundgeräusches, so daß die durch die Signalverarbeitung entstehende Reststörungen maskiert werden und das reduzierte Signal natürlicher klingt. Der Exponent γ, mit

$$\gamma > 0 , \tag{6.61}$$

bietet die Möglichkeit, die Schärfe des Schätzalgorithmus zu verändern und verschiedene Arten der Geräuschreduktion zu realisieren. Dabei ergibt sich für $\gamma = 0{,}5$ die Subtraktion der Betragsspektren und für $\gamma = 1$ die Leistungssubtraktion. Grundsätzlich zeigt sich aber, daß es besonders bei niedrigem SNR sehr schwierig ist, gleichzeitig eine hohe Störgeräuschunterdrückung und eine minimale Nutzsignalverzerrung zu erzielen.

Durch adaptive Einstellung der Koeffizienten a, b und γ in Abhängigkeit vom Verlauf der globalen Mithörschwelle $L_T(n, m)$ wird die Verbesserung der Geräuschreduktion bzgl. Verzerrung des Sprachsignals und psychoakustischer Verdeckung von Reststörgeräuschen erreicht.

6.5.2 Optimale Filterfunktion

Bereits in [13], [14], [72] und [151] wurden Verfahren vorgestellt, die durch Formulierung einer psychoakustisch motivierten Gewichtsfunktion ermöglichen, Reststörgeräusche und Verzerrungen des Sprachsignals durch Verdeckungseffekte zu verringern. Diese Verfahren bilden die Ausgangsbasis für die Bestimmung der optimalen Filterfunktion $G_{\min}(n, m)$. Sie wurden für den Einsatz im vorliegenden Geräuschreduktionssystem angepaßt und erweitert. Sofern Störung $n(k)$ und Nutzsignal $s(k)$ orthogonal zueinander sind, folgt für das spektrale Fehlersignal zwischen geschätztem und wirklichem Signalspektrum:

$$E(n,m) = \hat{S}(n,m) - S(n,m). \tag{6.62}$$

Mit (6.57) wird die spektrale Leistungsdichte $R_E(n, m)$ der Signalverzerrung $E(n, m)$ zu:

$$\begin{aligned} R_E(n,m) &= R_{ss}(n,m)\cdot\left[G_W(n,m)-1\right]^2 + R_{nn}(n,m)\cdot\left[b - a\cdot G_W(n,m)\right]^2 \\ &= R_{Es}(n,m) + R_{En}(n,m). \end{aligned} \tag{6.63}$$

Das LDS des Fehlersignals $R_E(n, m)$ beinhaltet damit die Fehlerkomponenten $R_{Es}(n, m)$ und $R_{En}(n, m)$, die von den Verzerrungen des Nutzsignals und der Reststörung herrühren.

Es gibt nun verschiedene Möglichkeiten, eine psychoakustische Gewichtsfunktion $G_s(n, m), G_n(n, m)$ oder $G_{\min}(n, m)$ in Abhängigkeit von der globalen Maskierschwelle $R_T(n, n)$ zu finden: Werden zunächst Verzerrungen $R_{Es}(n, m)$ des Sprachsignals verdeckt, gilt gemäß (6.63):

$$R_T(n,m) = R_{ss}(n,m)\cdot\left[G_s(n,m)-1\right]^2. \tag{6.64}$$

Durch Umstellung nach $G_s(n, m)$ erhält man die Berechnungsvorschrift für ein adaptives Filter, das alle Verfälschungen $R_{Es}(n, m)$ des Sprachsignals verdeckt.

Es gilt dann nämlich:

$$G_s(n,m) = 1 - \sqrt{\frac{R_T(n,m)}{R_{ss}(n,m)}}, \tag{6.65}$$

da $0 \le G_s(n,m) \le 1$ erfüllt sein muß. Die Wahl einer derartigen Gewichtsfunktion $G_s(n,m)$ für die psychoakustische Geräuschreduktion hat folgende Konsequenzen: Ist $R_{nn}(n,m) \gg R_{ss}(n,m)$, d.h. das SNR $\xi(n,m)$ ist klein, dann bleiben folgende Verzerrungen hörbar und man erhält den Fehler:

$$R_E = R_{En}\left(G_s(n,m)\right) - R_T(n,m). \tag{6.66}$$

Dadurch werden im Vergleich zu $G_{min}(n,m)$ gemäß (6.72) die Verzerrungen des Sprachsignals vermindert, doch die Verzerrung der Reststörgeräusche nehmen zu.

Gilt dagegen $R_{nn}(n,m) \ll R_{ss}(n,m)$, dann ist auch die Maskierschwelle $R_T(n,m)$ sehr hoch. Hier ergeben sich stärkere Verzerrungen des Sprachsignals verglichen mit (6.72). Das Verfahren arbeitet in diesem Fall nicht effektiv. In Abbildung 6.22 werden diese Effekte verdeutlicht. Optimiert man die Geräuschreduktion hinsichtlich der primären Verdeckung der Störsignalverzerrungen (z.B. „musical tones“) und nimmt dafür mehr hörbare Verzerrungen des Sprachsignals in Kauf, so gilt:

$$R_T(n,m) = R_{nn}(n,m) \cdot \left[b - a \cdot G_n(n,m)\right]^2. \tag{6.67}$$

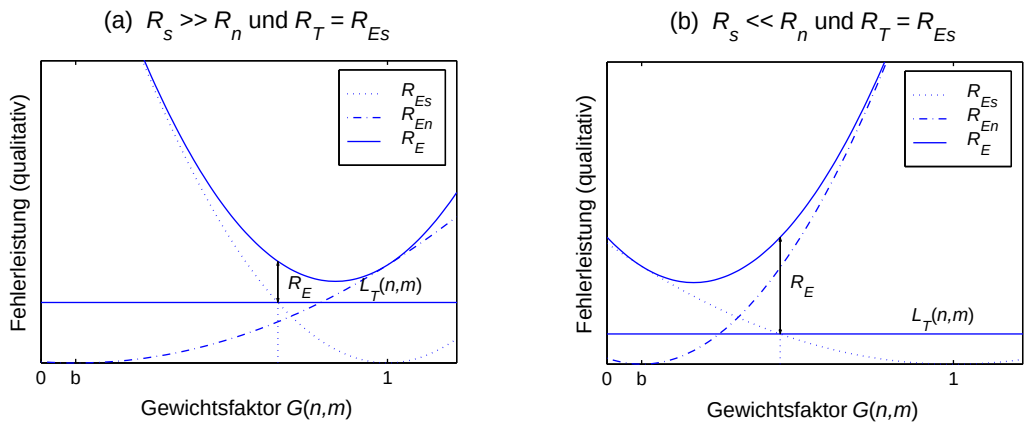


Abbildung 6.22 Gewichtsfunktion zur primären Verdeckung von Verzerrungen des Nutzsignals. Die wahrnehmbaren Anteile sind eingezeichnet. Qualitative Darstellung.

Durch Umstellung von Gleichung (6.67) erhält man die psychoakustische Gewichtsfunktion $G_n(n, m)$ aus

$$G_n(n,m) = \sqrt{\frac{R_T(n,m)}{a^2 \cdot R_{nn}(n,m)}} + b, \tag{6.68}$$

wobei gilt $0 \le G_n(n, m) \le 1$. Diese Verfahrensweise hat zwei Konsequenzen: Bei kleinem SNR $\xi(n, m)$ nehmen im Vergleich mit (6.72) zwar die Reststörgeräusche zu, dagegen werden hörbare Verzerrungen des Nutzsignals reduziert. Das Nutzsignal verbessert sich um:

$$R_E = R_{Es}\left(G_n(n,m)\right) - R_T(n,m). \tag{6.69}$$

Für großes SNR ergibt sich eine relativ hohe Maskierschwelle $R_T(n, m)$, wodurch oft alle Reststörgeräusche bereits verdeckt werden. Verglichen mit (6.72) werden zusätzlich die Verzerrungen des Nutzsignals minimal gehalten. Liegt das gesamte Störsignal unterhalb der Maskierschwelle, kann die Geräuschreduktion deaktiviert werden. Beide Szenarien sind in Abbildung 6.23 schematisch dargestellt.

Durch die Optimierung der Gewichtsfunktion hinsichtlich der primären Verdeckung der unnatürlichen Störgeräusche werden die Reststörgeräusche so „geformt“, daß sie entweder verdeckt werden oder angenehme Reststörungen beinhalten, die sich nicht negativ auf den Höreindruck auswirken.

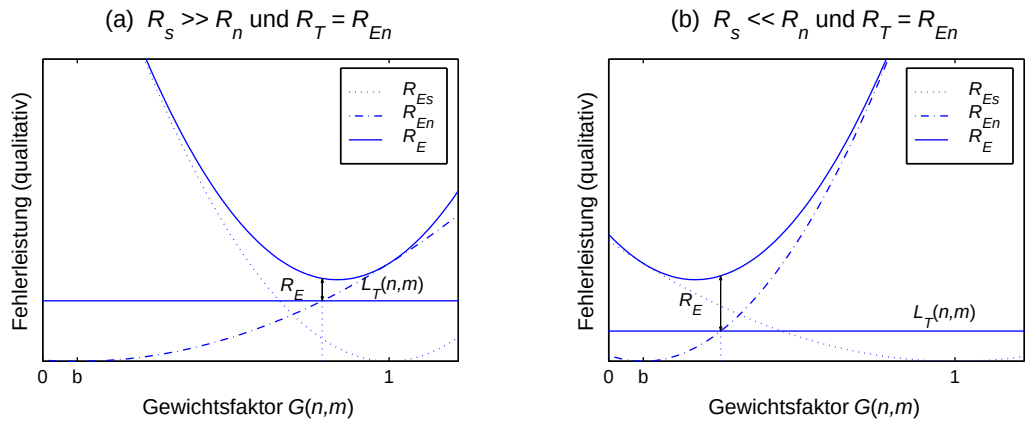


Abbildung 6.23 Gewichtsfunktion zur primären Verdeckung von Reststörgeräuschen. Die wahrnehmbaren Anteile sind eingezeichnet. Qualitative Darstellung.

6.5.3 Einstellung der Parameter

Die Bestimmung der Parameter a und b wird am Gesamtsystem vorgenommen. Dabei wurde auf das bereits behandelte CA-Verfahren zur Schätzung der Stör- und Nutzsignalleistungsdichte zurückgegriffen. Liegt die Leistungsdichte des Reduktionsfehlers $R_E(n, m)$ gemäß Gleichung (6.63) im *m*-ten Zeitfenster und einer bestimmten Frequenz n unterhalb der Maskierschwelle $R_T(n, m)$, so ist dieser Fehler absolut nicht wahrnehmbar. Dies gilt sowohl für die Verzerrungen des Sprachsignals $R_{Es}(n, m)$ wie auch für die Verfärbungen des Reststörgeräusches $R_{En}(n, m)$. Durch Differentiation von Gleichung (6.63) gemäß:

$$\frac{\partial R_E(n,m)}{\partial G(n,m)} = 0 \tag{6.70}$$

und wegen

$$\frac{\partial^2 R_E(n,m)}{\left[\partial G(n,m)\right]^2} > 0 \tag{6.71}$$

ergibt sich die *minimal* erreichbare Fehlerleistung $R_{Emin}(n, m)$ mit der dazugehörigen Gewichtsfunktion $G_{min}(n, m)$:

$$\begin{aligned} G_{\min}(n,m) &= \frac{R_{ss}(n,m) + abR_{nn}(n,m)}{R_{ss}(n,m) + a^2 R_{nn}(n,m)} \\ &= \frac{\xi(n,m) + ab}{\xi(n,m) + a^2}. \end{aligned} \tag{6.72}$$

Dabei gilt für die entsprechende Gewichtfunktion $0 \le G_{min}(n, m) \le 1$. In Abbildung 6.25 ist das Leistungsdichtespektrum des Fehlersignals $R_E(n, m)$ qualitativ dargestellt. Es zeigt sich, daß im allgemeinen Fall nicht davon ausgegangen werden kann, daß die minimale Fehlerleistungsdichte $R_E(n, m)$ unterhalb der Maskierschwelle $R_T(n, m)$ liegt. Die durch die Geräuschreduktion entstehenden Verzerrungen werden somit nicht völlig verdeckt und bleiben hörbar. Damit das gesamte Fehlersignal nicht hörbar ist muß aber gelten

$$R_T(n,m) \ge R_{E\min}(n,m). \tag{6.73}$$

Durch Einsetzen von (6.72) in (6.63) und Beachtung der Bedingung aus (6.73) ergibt sich mit zum Beispiel $a = 1$ der Parameter b, wobei:

$$1 \geq b \geq 1 - \sqrt{R_T(n,m) \cdot \frac{R_{xx}^2(n,m)}{R_{ss}(n,m)R_{nn}^2(n,m) + R_{nn}(n,m)R_{ss}^2(n,m)}} \geq 0. \tag{6.74}$$

Mit der Anhebung des *Spectral Floors* b läßt sich die psychoakustisch optimale Filterfunktion $G_{min}(n,m)$ so bestimmen, daß sowohl alle Verzerrungen des Sprachsignals wie auch alle Reststörgeräusche völlig maskiert werden. Genau genommen entspricht dies der Verdeckung der Fehlersignale $R_{Es}(n,m)$ und $R_{En}(n,m)$ durch die bewußte Verringerung der Dämpfung des Störgeräusches $n(k)$. Hörtests haben gezeigt, daß Verfahren mit einem derart hinzugefügtem Störgeräusch solchen Verfahren, die eine völlige Reduktion der Reststörungen bewirken, bevorzugt werden. Außerdem ist nun für den fernen Teilnehmer einer Freisprechverbindung in Sprachpausen besser erkennbar, ob eine Telefonverbindung in das Fahrzeug noch besteht. Ohne verbleibende Fahrgeräusche im übertragenen Freisprechsignal wäre zudem die Lautstärke und Anstrengung der Sprecherstimme im Fahrzeug für den fernen Gesprächsteilnehmer schwer nachvollziehbar. Es hat sich gezeigt, daß der Parameter b die psychoakustische Verdeckung von Reduktionsfehlern des Nutzsignals maßgeblich beeinflußt. Erhöht man b, so wird dem geräuschreduziertem Signal $\hat{s}(k)$ eine Reststörung überlagert. Das hat den Effekt, daß zwar die resultierende SNR-Verbesserung des Algorithmus abnimmt, aber sich dafür der Gesamthöreindruck des geräuschreduzierten Signals verbessert. Stellt man als Sonderfall der Gleichung (6.75) $b = 0$ ein, so können im allgemeinen nicht alle Verzerrungen des Sprachsignals und der Reststörung verdeckt werden. Sie bleiben damit hörbar. Diesen Umstand sieht man deutlich in Abbildung 6.24.

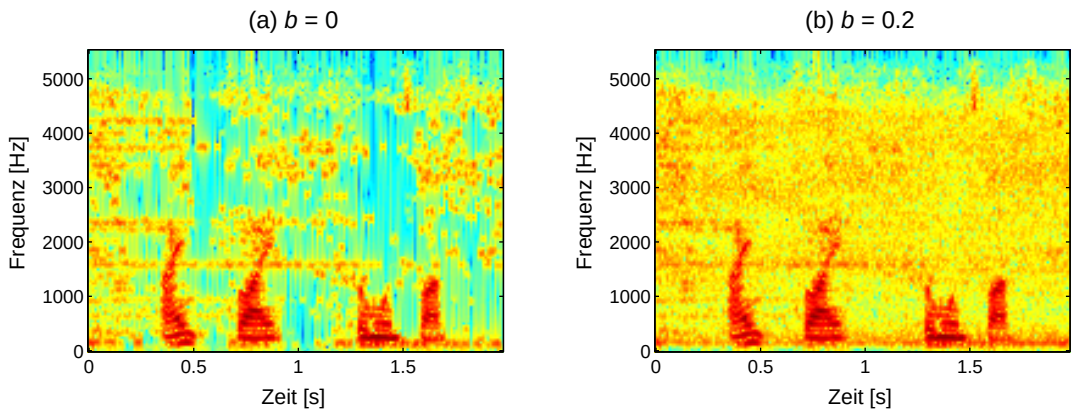


Abbildung 6.24 Einfluß des Parameters b auf das *CA-LPC*-Verfahren. **(a)** Deutliche Artefakte und Reststörungen für $b = 0$. **(b)** Störungen sind im Spectral Floor mit $b = 0.2$ nicht mehr hörbar.

Generell zeigt sich, daß Verfärbungen des Restgeräusches als störender empfunden werden als kleine Verzerrungen des Sprachsignals. Daher macht es durchaus Sinn, einen vorgegebenen „spectral floor“ b einzustellen, um damit Verfärbungen der Reststörung psychoakustisch zu verdecken.

Hörtests haben diesen bereits theoretisch beschriebenen Fakt für $0.05 \le b \le 0.4$ bestätigt. Für alle hier vorgestellten psychoakustischen Verfahren wurde der Wert $b = 0.2$ gewählt und unveränderlich eingestellt.

Durch die Modifikation der Gewichtsfunktion (6.72) anhand des Oversubtraction-Faktors a kann jetzt eine völlige Fehlerverdeckung erreicht werden. Dies ist in Abbildung 6.25 sichtbar. Die dargestellten Parabeln werden durch die Veränderung von a unterschiedlich gespreizt. Setzt man neben Gleichung (6.73) weiter voraus, daß b fest eingestellt wird, dann kann gemäß Gleichung (6.63) nur eine völlige Maskierung aller Fehler erreicht werden, wenn der Oversubtraction-Faktor a adaptiv zu:

$$0 \le \frac{b \cdot R_{nn} \cdot R_{ss} - \sqrt{R_{nn} \cdot R_{ss} \cdot R_T \cdot \left(R_{ss} + b^2 \cdot R_{nn} - R_T\right)}}{R_{nn} \cdot \left(R_{ss} - R_T\right)} \le a \le 1 \tag{6.75}$$

gewählt wird. Zur Verbesserung der Lesbarkeit wurden in Gleichung (6.75) die Zeit- und Frequenzindizes n, m weggelassen. Um den Verlauf von $a(n, m)$ zu glätten, wurde eine anschließende Medianfilterung fünften Grades durchgeführt.

.

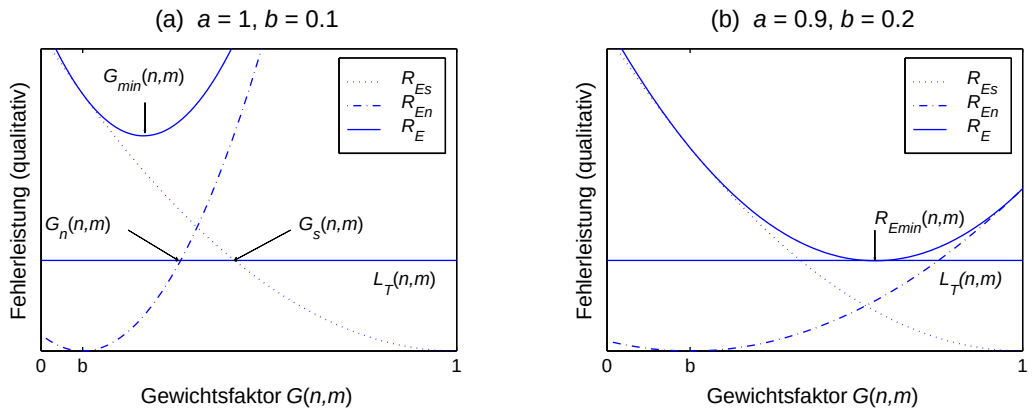


Abbildung 6.25 Reduktionsfehlersignal $R_E(n, m)$ und seine Komponenten $R_{Es}(n, m)$ bzw. $R_{En}(n, m)$. **(a)** Die Spreizung der Parabeln ist von der Leistungsdichte des Nutzsignals und der Störleistung sowie vom Oversubtraction-Faktor a abhängig. In **(b)** wird durch Änderung des Spectral Floors b und des Oversubtraction-Faktors a gerade $R_E(n, m) = R_T$ erreicht. Qualitative Darstellung.

Generell wäre auch die feste Einstellung von a denkbar. Ein derartiger Ansatz wurde zum Beispiel in [72] mit $a = 1$ gewählt. Dort wurde im vorgestellten Verfahren auf die Verdeckung der Verzerrungen $R_{En}(n, m)$ des Störgeräusches, die als besonders störend empfunden werden, Wert gelegt. Dadurch kommt es aber zu einer geringen Verschiebung des Wienerfilters bezüglich des minimalen Gesamtfehlers und zur zusätzlichen Verzerrung des Sprachsignals.

Besonders bei großem SNR zeigen sich dagegen die Vorteile des in dieser Arbeit vorgestellten Verfahrens mit adaptiver Einstellung des Oversubtraction-Faktors a. Es wird ein optimaler Kompromiß zwischen den unvermeidbaren Verzerrungen des Sprachsignals $R_{Es}(n, m)$ und den Verfärbungen des Störsignals $R_{En}(n, m)$ gefunden, wobei der Gesamtfehler $R_E(n, m) = R_{Es}(n, m) + R_{En}(n, m)$ durch die adaptive Modifikation der Filterregel gemäß den Gleichungen (6.57) und (6.72) völlig verdeckt wird. Dies wird beispielhaft in Abbildung 6.25 verdeutlicht. Das neue Verfahren zeichnet sich durch ein nahezu unverfärbtes Reststörgeräusch und durch vergleichbar geringe Verzerrungen des Sprachsignals bei dennoch hoher Geräuschreduktion und Sprachverständlichkeit aus. Hör- und instrumentelle Tests haben diese Aussagen bestätigt. Abbildung 6.26 zeigt den Verlauf des geglätteten Oversubtraction-Faktors a für eine im Fahrzeug aufgenommene gestörte Sprachprobe für $b = 0.2$.

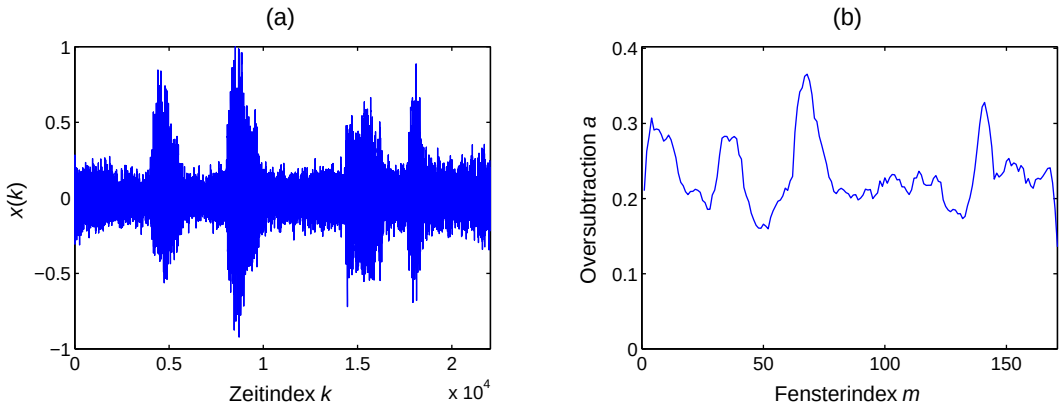


Abbildung 6.26 Adaptive Einstellung des Oversubtraction-Faktors a. **(a)** Gestörtes Testsignal $x(k)$, aufgenommen im BMW 528i touring bei Regen und 120 km/h. **(b)** adaptiv eingestellter Faktor a.

6.6 Validierung und Bewertung des Gesamtsystems

Ausgehend von dem in Abbildung 6.1 gezeigten Blockdiagramm für die Realisierung einer psychoakustischen einkanaligen Geräuschreduktion werden folgende Teilschritte und Berechnungen in dem entwickelten Gesamtsystem durchgeführt:

- Aufnahme des gestörten Mikrophonsignals $x(t)$, Abtastung und Quantisierung zum diskreten Eingangssignal $x(k)$

- gewichtete Fensterung und Überlappung des Eingangssignals zu $x(k, m)$
- diskrete Fouriertransformation und Berechnung des Kurzzeitleistungsdichte $\tilde{R}_{xx}(n, m)$ des gestörten Sprachsignals
- Schätzung der Geräuschleistungsdichte $\hat{R}_{nn}(n, m)$ mit dem Pausen-, Min- oder CA-Verfahren
- Schätzung der Nutzsignalleistungsdichte $\hat{R}_{ss}(n, m)$ und Formantenrekonstruktion mit dem CA-, LPC- oder CC-Verfahren
- Bestimmung der lokalen psychoakustischen Mithörschwellen anhand lokaler Maskiermodelle und der geschätzten Sprachsignalleistungsdichte, nichtlineare Superposition und Bestimmung der resultierenden globalen Mithörschwelle $R_T(n, m)$
- Bestimmung der psychoakustischen Gewichtungsregel und Parametrisierung der spektralen Subtraktion, adaptive Bestimmung der Parameter a und b, Minimierung der hörbaren Sprachverzerrung und Verfärbung des Geräusches
- Parametrische Filterung des Eingangssignalspektrums $X(n, m)$ und Bestimmung des gefilterten Ausgangsspektrums $\hat{S}(n, m)$
- Overlap-and-Add-Methode und Rücktransformation in den Zeitbereich für das zeitdiskrete, geräuschreduzierte Ausgangsignal $\hat{s}(k)$.

Diese Aufstellung macht deutlich, daß verschiedene Faktoren und Fehlerquellen großen Einfluß auf die Qualität des resultierenden Ausgangssignals $\hat{s}(k)$ haben.

Als besonderes kritisch werden die Schätzung der Störleistungsdichte und die Bestimmung der psychoakustischen Mithörschwelle, die Rekonstruktion der Formantstruktur und letztendlich die Parametrisierung der psychoakustischen Gewichtungsfunktion eingeschätzt. Deshalb wurden diese Schwerpunkte bereits sehr ausführlich behandelt. Für jeden dieser bedeutenden Teilschritte wurden Alternativen und Lösungsvorschläge untersucht. In den nachfolgenden Betrachtungen werden nur diese Methoden und Verfahren hinsichtlich der Gesamtbewertung variiert und gemäß Tabelle 6.2 miteinander verglichen.

Beim *NL-SS*-Verfahren erfolgt keine psychoakustische Gewichtung. Dafür wird eine nichtlineare Nachfilterung (hier: Medianfilterung) vorgenommen. Außerdem wurden alle Verfahren mit der einfachen Spektralen Subtraktion (*CA-SS*) verglichen. Der Vergleich erfolgt anhand subjektiver und instrumenteller Bewertungsmethoden aus Abschnitt 5.3.

Ein Vergleich zu mehrkanaligen Ansätzen wurde wegen der prinzipiellen Unterschiede der Verfahren nicht durchgeführt.

Bezeichnung	Schätzung LDS-Störung	Bestimmung Mithörschwelle	Gewichtung
CA-CA	CA-Verfahren	CA-Verfahren	psychoakustisch, adaptive Einstell. von a
CA-LPC	CA-Verfahren	LPC-Schätzung	psychoakustisch, adaptive Einstell. von a
CA-CC	CA-Verfahren	CC-Verfahren	psychoakustisch, adaptive Einstell. von a
MIN-MIN	Min-Verfahren	Min-Verfahren	psychoakustisch
NL-SS	CA-Verfahren	-	nichtlineare Spektrale Subtraktion
CA-SS	CA-Verfahren	-	Spektrale Subtraktion
ohne	keine	keine	ohne Verarbeitung

Tabelle 6.2 Testkonfiguration und Kombination verschiedener Verfahren

6.6.1 Instrumentelle Bewertung

Zunächst erfolgt die Bewertung mit instrumentellen Methoden. Dabei werden erneut verschiedene Geräuschszenarien berücksichtigt. Abbildung 6.27 zeigt die Bewertung unterschiedlicher Verfahren laut Tabelle 6.2 anhand der LPC-Distanz zwischen ungestörtem Sprachsignal $s(k)$ und geräuschreduziertem Signal $\hat{s}(k)$, siehe dazu auch Abschnitt 5.3.2. Die LPC-Distanz erfaßt dabei vor allem die Verzerrungen des Sprachsignals bezüglich seiner durch den Sprecherzeugungsprozeß aufgeprägten Eigenschaften. Die akustische Wahrnehmung des geräuschreduzierten Sprachsignals wird hierbei relativ wenig berücksichtigt.

Die Formung der Laute im Vokaltrakt läßt sich vornehmlich als autoregressiver Prozeß modellieren, siehe dazu Abschnitt 3.1.1. Die LPC-Analyse und Distanzbetrachtung ergibt bei so modellierbaren Prozessen zuverlässige Ergebnisse. Es zeigt sich, daß alle Verfahren eine relativ gute Rekonstruktion der Formantstruktur und der Vokale im Sprachsignal ermöglichen. Lediglich das Minimum-Verfahren weist hier bei großem SNR einen deutlichen Unterschied zu anderen Verfahren auf. Dies läßt sich auf die Art der Geräuschschätzung zurückführen, mit der die Formantstruktur nur unzureichend nachgebildet werden kann, vgl. Abschnitt 6.2.2. Auffällig ist, daß die Verfahren *CA-SS* und *NL-SS* bei rosa Rauschen und bei Regenfahrt mit mittleren Signal-Stör-Verhältnissen schlechtere Ergebnisse bei der Signalrekonstruktion liefern als die restlichen Verfahren.

Gerade bei rauschartigen Störern, die nicht mit autoregressiven Prozessen modelliert werden können, zeigen sich die Vorzüge der *LPC*- und *CC*-Verfahren für die Schätzung der Nutzsignale in gestörter Umgebung.

Erfreulicherweise schneidet das hier entwickelte und vorgestellte *CA-CA*-Verfahren sehr gut bei der Signalrekonstruktion ab. Es ist in nahezu allen Signal-Stör-Konstellationen den anderen Verfahren bei der Nachbildung der tonalen Struktur der Sprache überlegen. Ähnliche Ergebnisse liefert auch die Bewertung anhand der kepstralen Distanz zwischen Sprachsignal und geschätztem geräuschreduziertem Signal in Abbildung 6.28.

Abbildung 6.27 und Abbildung 6.28 zeigen den Vergleich verschiedener Geräuschreduktionsverfahren anhand von Systemdistanzmaßen, die aus der Modellierung des Spracherzeugungsprozesses hervorgehen. Nun erfolgt der Vergleich mit der sogenannten Barkdistanz, siehe Abschnitt 5.3.3. Die Barkdistanz wurde aus der Modellbildung psychologischer und physiologischer Eigenschaften des Gehörs gewonnen. Sie gibt darüber Aufschluß, wie die unterschiedlichen Verfahren der Geräuschreduktion wahrgenommen werden und welchen Höreindruck sie hinterlassen.

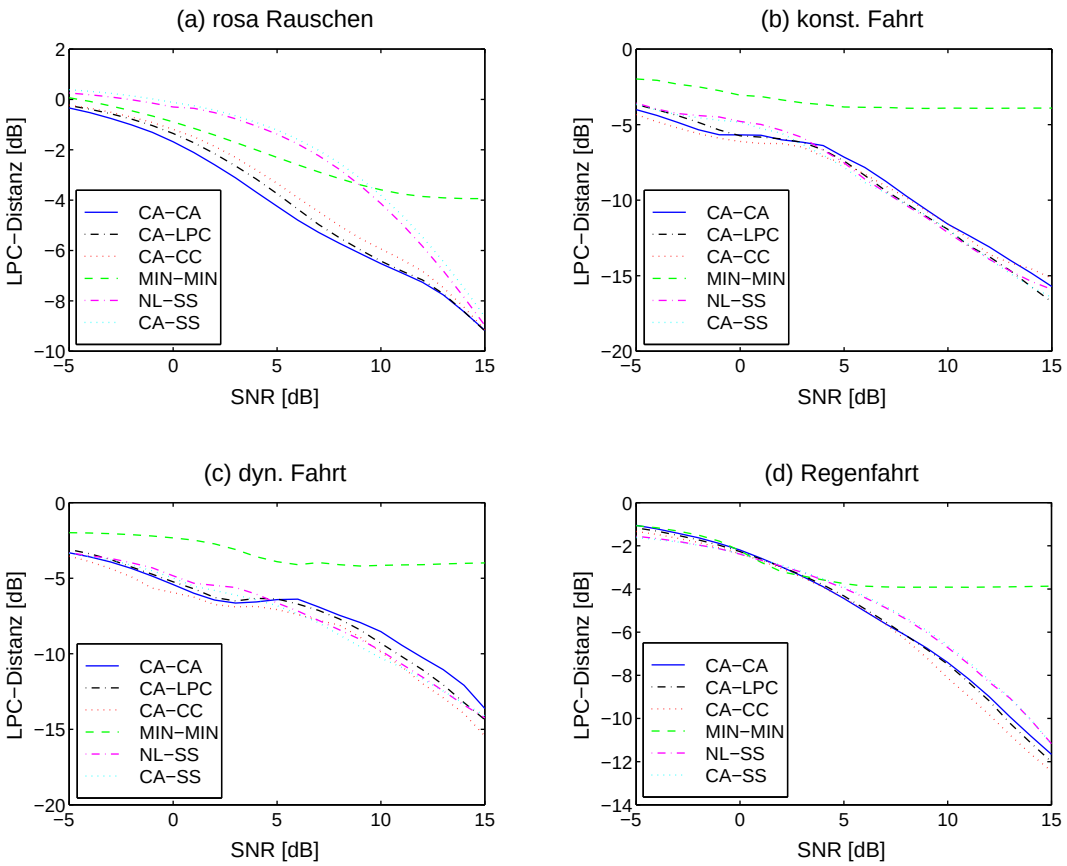


Abbildung 6.27 Vergleich unterschiedlicher Verfahren zur einkanaligen Geräuschreduktion. Bewertung mit der LPC-Distanz zwischen ungestörtem Sprachsignal $s(k)$ und geräuschreduziertem Signal $\hat{s}(k)$.

Genau genommen ist die Barkdistanz ein Leistungsabstandsmaß. Dabei werden die einzelnen Anregungen des ungestörten Sprachsignals in sogenannten Frequenzgruppen zusammengefaßt und mit den Frequenzgruppenleistungen, den Intensitäten des geräuschreduzierten Signals, verglichen. Somit kommen die eingangs erwähnten psychoakustischen Verdeckungseffekte zum Tragen. Je kleiner die Barkdistanz ist, desto ähnlicher ist die Wahrnehmung des geschätzten Sprachsignals $\hat{s}(k)$ zum geräuschfreien Sprachsignal $s(k)$ und desto besser klingt das geräuschreduzierte Signal. Abbildung 6.30 zeigt die Ergebnisse für verschiedene Verfahren während Sprachaktivität. Abbildung 6.31 zeigt den Vergleich in Sprachpausen.

Gerade in den Sprachpausen wirken sich Verzerrungen des Reststörgeräusches negativ auf den Gesamthöreindruck aus. Musical Tones oder andere Verfärbungen werden sofort wahrgenommen. Hier haben vor allem die Verfahren *CA-SS* und *NL-SS* Nachteile, da bei diesen keine psychoakustische Modifikation stattfindet. Erstaunlicherweise schneidet auch das *MIN-MIN*-Verfahren hier relativ schlecht ab. Dies läßt sich auf die nicht erwartungstreue Schätzung der Störleistungsdichte zurückführen.

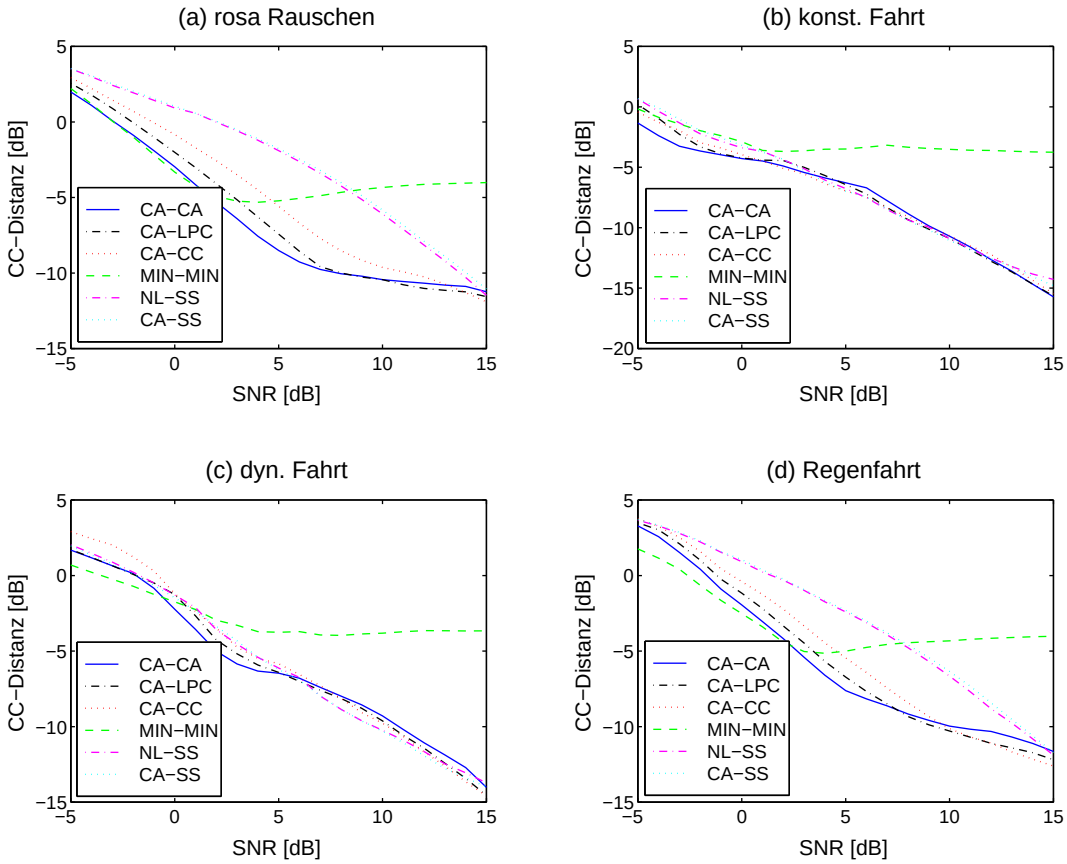


Abbildung 6.28 Vergleich verschiedener Verfahren zur einkanaligen Geräuschreduktion. Bewertung mit der kepstralen Distanz zwischen ungestörtem Sprachsignal $s(k)$ und geräuschreduziertem Signal $\hat{s}(k)$.

Besonders gute Ergebnisse sowohl in Sprachpausen wie auch während Sprachaktivität wurden mit den Verfahren *CA-LPC* und *CA-CC* erreicht. Hier zeigen sich vor allem die Vorteile, die sich durch die psychoakustische Parametrisierung der Verfahren ergeben. Innerhalb der betrachteten Frequenzgruppen werden die Verzerrungen des Sprachsignals und der Reststörungen psychoakustisch verdeckt.

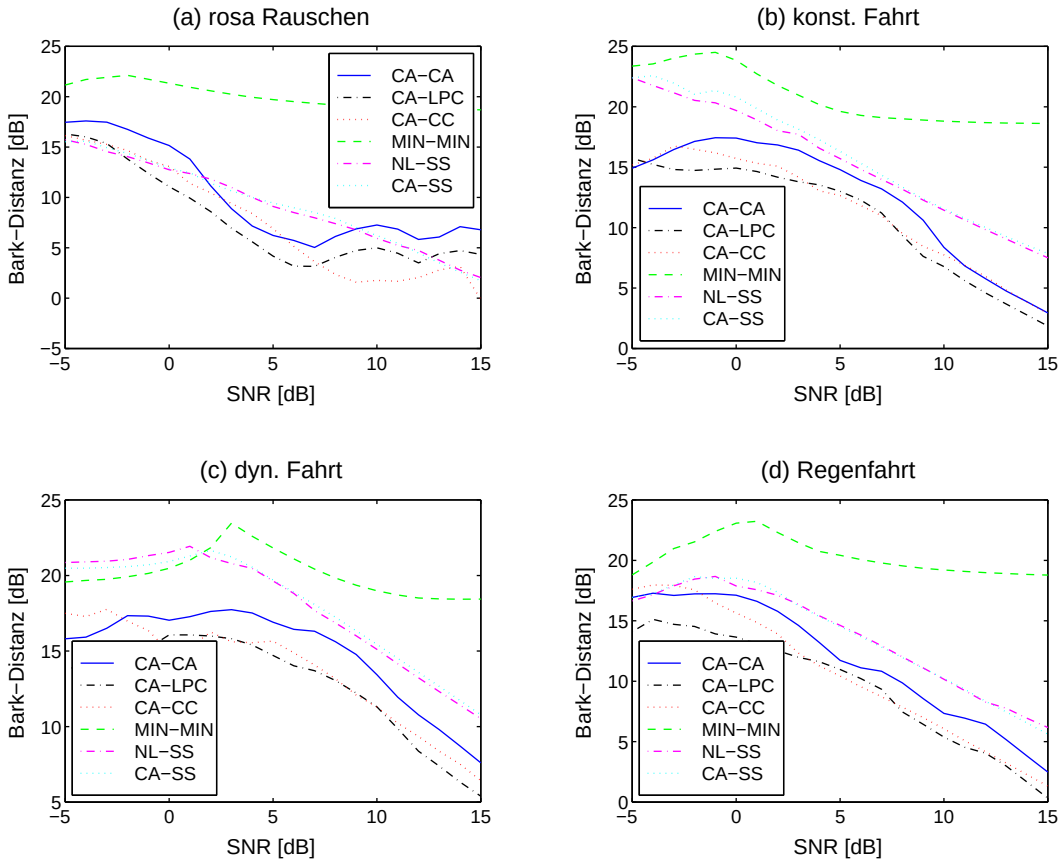


Abbildung 6.29 Vergleich unterschiedlicher Verfahren zur einkanaligen Geräuschreduktion. Bewertung mit der psychoakustischen Bark-Distanz zwischen ungestörtem Sprachsignal $s(k)$ und geräuschreduziertem Signal $\hat{s}(k)$.

6.6.2 Subjektive Bewertung

Leider gibt es bisher noch keine rein instrumentellen Maße oder Bewertungsmethoden zur Beurteilung der Sprachverständlichkeit. Die Verbesserung der Sprachverständlichkeit stellt das wichtigste Zielkriterium bei der Beurteilung der verschiedenen Verfahren dar. Deshalb wurden die unterschiedlichen Verfahren in umfangreichen Hörtests untersucht. Der Vergleich erfolgt mit subjektiven Bewertungsverfahren gemäß Abschnitt 5.2. Hier wurde ausschließlich der Mean-Opinion-Score (*MOS*) verwendet. Die *MOS*-Bestimmung erfolgt nach den Bewertungsmaßstäben in Tabelle 5.1. Die Ergebnisse für $SNR = 0\text{dB}$ sind in Tabelle 6.3 zusammengefaßt. Tabelle 6.4 zeigt die Bewertung für $SNR = 5\text{dB}$.

Bedenkt man die besondere Gewichtung der Bewertung bezüglich Sprachverständlichkeit, so zeigt sich, daß für alle Geräuscharten gute Ergebisse mit den psychoakustischen Verfahren erreicht werden. Vor allem die *CA-LPC* und *CA-CA*-Verfahren erzielen die beste Bewertung in den einzelnen Signal-Geräusch-Konstellationen. Besonders auffällig ist, daß das *NL-SS* und *CA-SS* in einigen Disziplinen schlechtere Bewertungen erhalten, als das eigentliche gestörte Eingangssignal. Das liegt vor allem an den teilweise sehr starken Verzerrungen des Sprachsignals und der Reststörung (musical tones). Dieser Effekt tritt vor allem bei kleinem SNR auf. Insgesamt wird deutlich, daß die psychoakustischen Verfahren besonders bei rauschähnlichen Störern (rosa Rauschen, Regenfahrt) erstaunlich gute Ergebisse liefern und die Sprachverständlichkeit deutlich verbessern. Breitbandiges Rauschen erzeugt innerhalb der einzelnen Frequenzgruppen eine deutlich höhere lokale Intensität als tonale Signale, was zur Simultanverdeckung führt. Die Ursache dafür wurde im Abschnitt 3.3.1.1 erklärt. Demnach erzielen psychoakustische Verfahren bei solchen Störern besonders gute Ergebnisse. Sie nutzen Maskiereffekte, um etwaige Verzerrungen des Sprachsignals und der Reststörungen wirkungsvoll zu verdecken.

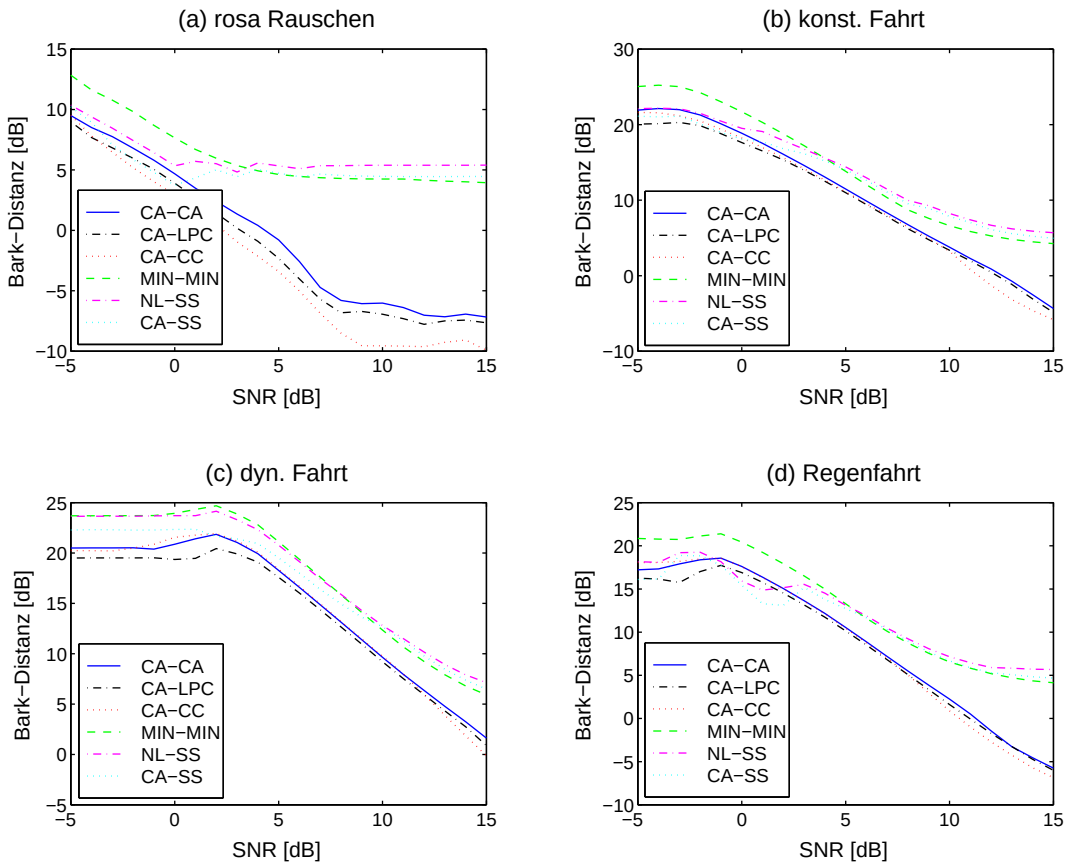


Abbildung 6.30 Vergleich unterschiedlicher Verfahren zur einkanaligen Geräuschreduktion. Bewertung mit der psychoakustischen Bark-Distanz zwischen gestörtem Sprachsignal $x(k)$ und Reststörgeräusch $\hat{n}(k)$ in Sprachpausen.

Die Verfahren *NL-SS* und *CA-SS* waren dagegen allen anderen Verfahren bezüglich der Sprachverständlichkeit und des insgesamten Höreindrucks unterlegen. Vergleicht man die subjektiven Ergebnisse aus Tabelle 6.3 und Tabelle 6.4 mit den instrumentellen Ergebnissen in Abbildung 6.29, so zeigt sich ein ähnliches qualitatives Bild. Dennoch wurde von den Testpersonen das *MIN-MIN*-Verfahren subjektiv besser empfunden, als dies die instrumentellen Ergebnisse vermuten ließen. Von der Vielzahl der Testpersonen wurden das *CA-CA*- und *CA-LPC*-Verfahren anderen Verfahren vorgezogen. Es wurde zum Teil die leichte Hochpaßcharakteristik dieser Verfahren bemängelt, die aber keinen negativen Einfluß auf die Verständlichkeit des geschätzten Sprachsignals hatte.

Verfahren	**rosa Rauschen**	**konstante Fahrt**	**dynamische Fahrt**	**Regenfahrt**
CA-CA	2.1	**2.0**	**3.0**	**3.0**
CA-LPC	**2.0**	**2.0**	3.3	**3.0**
CA-CC	2.2	2.1	3.1	3.5
Min-Min	3.2	2.2	4.0	3.5
NL-SS	3.3	3.3	4.5	4.3
CA-SS	3.9	3.5	4.8	4.9
ohne	3.7	3.0	3.5	4.3

Tabelle 6.3 MOS-Bewertung der Verfahren für $SNR = 0\text{dB}$.

Verfahren	**rosa Rauschen**	**konstante Fahrt**	**dynamische Fahrt**	**Regenfahrt**
CA-CA	**1.7**	**2.1**	**2.4**	**2.3**
CA-LPC	1.8	**2.1**	2.5	**2.3**
CA-CC	1.8	2.2	2.6	2.5
Min-Min	2.5	2.8	3.0	3.0
NL-SS	2.9	2.2	2.8	3.0
CA-SS	3.3	2.6	3.3	3.5
ohne	3.0	2.5	2.8	3.6

Tabelle 6.4 MOS-Bewertung der Verfahren für $SNR = 5\text{dB}$.

Fast man sowohl die instrumentellen, wie auch die subjektiven und makroskopischen Untersuchen zusammen, ergeben sich die besten Bewertungen für das *CA-CA* oder *CA-LPC*-Verfahren. Ungefähr 75% Prozent aller beteiligten Testhörer haben den psychoakustischen Verfahren zur Signalverbesserung eine bessere Sprachverständlichkeit bescheinigt.

6.6.3 Makroskopischer Vergleich

Abbildung 6.31 zeigt die Barkgramme des gestörten Sprachsignals $x(k)$ und des mit den einzelnen Verfahren geräuschreduzierten Signals $\hat{s}(k)$. Dabei wurden die Beispielaufnahmen bei starkem Regen im BMW 528i touring bei 70 km/h durchgeführt und mit verschiedenen Verfahren bearbeitet. Es ergab sich ein Eingangssignal-Stör-Abstand von $SNR = 5\,\text{dB}$.

Die Geräuschreduktion wurde bei allen Verfahren mit dem fest eingestellten spectral floor $b = 0{,}2$ vorgenommen. Bei den Verfahren *CA-CA* und *CA-LPC* wurde der Oversubtraction-Faktor a adaptiv eingestellt, während beim *CA-SS*, *NL-SS* und *MIN-MIN*-Verfahren $a = 0{,}95$ fest vorgegeben wurde.

Zunächst wird sofort sichtbar, daß das *CA-SS*-Verfahren die meisten Artefakte und sporadisch auftretende Reststörungen (musical tones) aufweist. Durch die Medianfilterung beim *NL-SS*-Verfahren wurden diese „verschmiert“ und damit deutlich reduziert. Vor allem im mittleren und unteren Barkbereich treten aber noch Störungen auf, die sich negativ auf den Gesamthöreindruck auswirken.

Das *MIN-MIN*-Verfahren nutzt bereits psychoakustische Aspekte der Geräuschreduktion. Dennoch sind Störungen im unteren Barkbereich wahrnehmbar. Es kommt teilweise zum „Blubbern“. Diese Effekte sind durch die besondere Art der Geräusch- und Maskierschwellenschätzung bedingt. Im mittleren und hohen Barkbereich hat dagegen die Geräuschreduktion sehr gute Ergebnisse erreicht. Dennoch sind gerade in Fahrgeräuschen besonders hohe Leistungsdichten im unteren Frequenzbereich vertreten, so daß sich hier Fehlschätzungen sofort negativ auf den Gesamteindruck auswirken.

Die besten Ergebnisse werden vom *CA-CA* und *CA-LPC*-Verfahren erreicht. Auch im unteren Barkbereich treten kaum Artefakte auf. Der „Einschwingvorgang“ bei der Adaption der Geräusch- und Maskierschwellenschätzung ist aber deutlich wahrnehmbar und sichtbar. Insgesamt ergibt sich bei diesen beiden Verfahren ein sehr guter Höreindruck, wobei beim *CA-CA*-Verfahren im höheren Barkbereich weniger Artefakte sichtbar sind.

In informellen Hörtests wurden aber nur von 15% der Testpersonen Unterschiede zwischen beiden Verfahren ausgemacht, wobei das *CA-CA*-Verfahren leicht bevorzugt wurde. Mit der makroskopischen Betrachtung sind nur Artefakte während der Sprachpausen identifizierbar. Aussagen zur Sprachverständlichkeit oder zu Verzerrungen des Sprachsignals können mit die-

ser Art der Bewertung nicht vorgenommen werden. Dazu sind umfangreiche Hörtests mit wechselnden Äußerungen und anschließender statistischer Auswertung notwendig. Dennoch vermitteln die in Abbildung 6.31 dargestellten Barkgramme einen guten Eindruck von den Ergebnissen der psychoakustischen Geräuschreduktion.

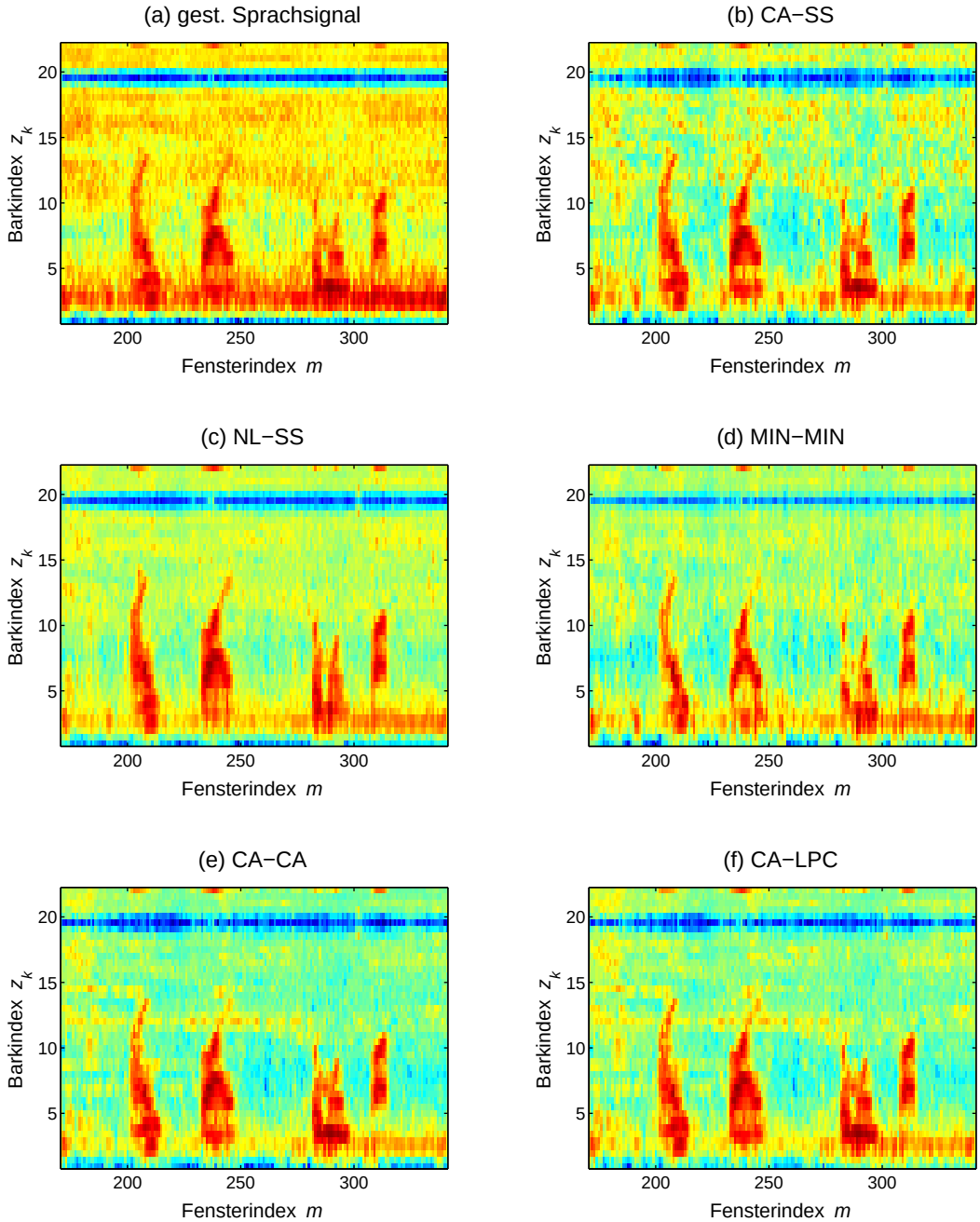


Abbildung 6.31 Verschiedene Geräuschreduktionsverfahren im Vergleich anhand der Barkgramme, vgl. Abschnitt 5.3.4. Aufnahme der Signale bei Regenfahrt im BMW 528i touring mit 130 km/h. **(a)** gestörtes Signal. Geräuschreduktion mit dem: **(b)** *CA-SS*-Verfahren **(c)** *NL-SS*-Verfahren **(d)** *MIN-MIN*-Verfahren **(e)** *CA-CA*-Verfahren **(f)** *CA-LPC*-Verfahren.

6.7 Komplexität und Echtzeitanforderungen

Alle Verfahren wurden auf einem PC mit Athlon® 1000 Mhz und mit MatLab® und DaDisp® simuliert und ausgewertet. Die Testaufnahmen wurden blockweise verarbeitet, so wie es auch in einem Echtzeitsystem realisiert werden würde. Für jedes einzelne Verfahren wurden die notwendigen Floating-Point-Operationen (Flops) gemessen. Die Ergebnisse sind für Testsignale von einer bzw. zwei Sekunden Länge in Abbildung 6.32 dargestellt.

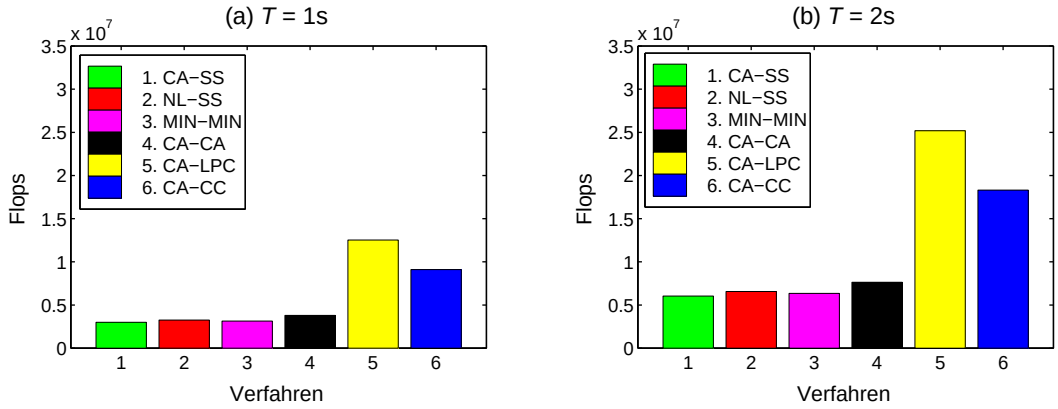


Abbildung 6.32 Vergleich der Komplexität der Verfahren in Floating-Point-Operations (Flops). (a) Signaldauer = 1s (b) Signaldauer = 2s.

Vor allem die Verfahren *CA-LPC* und *CA-CC* weisen eine hohe Komplexität und Rechenanforderung auf. Insgesamt ist die benötigte Rechenleistung mehr als doppelt so hoch wie bei den anderen Verfahren. Dies liegt vor allem an den sehr rechenintensiven Systemidentifikationen innerhalb der *LPC*- und CC-Analyse. Der Aufwand für die Geräuschschätzung und die Berechnung der Maskierschwellen unterscheidet sich dagegen nicht vom *CA-CA*-Verfahren. Vergleich man Aufwand und Ergebnis des *CA-CA*-Verfahrens mit z.B. dem *CA-LPC*-Verfahren, so ergibt sich eine ungünstige Leistungsbilanz für den *CA-LPC*- und *CA-CC*-Algorithmus. Hier werden DSP-Leistungen von bis zu 13 MFlops/s benötigt.

Bedenkt man jedoch, daß eine weitere Optimierung des Programmcodes eine geringere DSP-Last zur Folge hätte, sind bereits preiswerte aktuelle Floating-Point-DSP (z.B. TI-TMS320C67x) einsetzbar. Eine Umsetzung dieser Verfahren auf einem Fixed-Point-DSP erscheint nicht sinnvoll. Das *CA-CA*-Verfahren bietet dagegen die besseren Voraussetzungen bzgl. des Rechenaufwandes. Für dieses Verfahren sind äußerst preiswerte Festkomma-DSP (z.B. TI TMS320Cx) verwendbar. In Abbildung 6.32 wächst die Anzahl der Floating-Point-Operationen fast linear mit der Erhöhung der Dauer des zu verarbeitenden Signals. Damit sind Echtzeitanwendungen mit den genannten Leistungsvorgaben durchführbar.

7 Zusammenfassung und Ausblick

Geräuschreduktionsverfahren sind wegen der breiten Anwendung in Mobiltelefonen und vor allem in Kraftfahrzeugen bekannt. Dabei kommt aber der eigentlichen Reduktion von Störgeräuschen eine untergeordnete Rolle zu. Viel wichtiger ist die Verständlichkeit der Sprache in geräuschbehafteter Umgebung. Demnach wurde in dieser Arbeit nicht die Maximierung der Geräuschreduktion, sondern die Verbesserung der Sprachverständlichkeit in den Mittelpunkt der Untersuchungen gestellt.

Nach Vorstellung verschiedener ein- und mehrkanaliger Verfahren, die den Stand der Technik markieren, wurde detailliert auf neuere einkanalige Verfahren eingegangen, die vermehrt die Charakteristik des menschlichen Gehörs bei der Wahrnehmung und Beurteilung von Sprachsignalen als Ausgangspunkt haben. Hierbei kommt vor allem den sogenannten psychoakustischen Verdeckungseffekten im akustischen Wahrnehmungsprozeß eine sehr große Bedeutung zu. Aus der Modellierung des menschlichen Gehörs entstand ein Wahrnehmungsmodell, mit dem die eigentliche Geräuschreduktion parametrisiert werden kann, so daß eine hohe Geräuschreduktion mit minimaler Verzerrung des Sprachsignals erreicht wurde. Die Grundidee dieses neuen Geräuschreduktionsalgorithmus ist dabei, die bekannte parametrische Wienerfilterung so abzuwandeln, daß alle durch das Verfahren bedingten Verzerrungen des Sprachsignals und der Reststörungen unterhalb der sogenannten Mithörschwelle bleiben und damit nicht wahrnehmbar sind. Ausgehend von der parametrischen Filterung mit der Regel

$$\hat{S}(n,m) = S(n,m)\cdot G_W(n,m) + N(n,m)\cdot\left[b - a\cdot G_W(n,m)\right], \tag{7.1}$$

werden Sprachanteile $S(n,m)$ und Störanteile $N(n,m)$ selektiv gefiltert. Der dabei entstehende Fehler wird für das Nutzsignal und die Störung so minimiert, daß beide Anteile unterhalb der Mithörschwelle $R_T(n,m)$ liegen und damit psychoakustisch verdeckt werden. Besonders kritisch ist dabei die Auswahl der Parameter a und b. Der sogenannte spectral floor b wird bei den neuen Verfahren *CA-CA*, *CA-LPC* und *CA-CC* fest eingestellt, während der Oversubtraction-Faktor a adaptiv so gewählt wird, daß die Fehlerleistungsdichte der Filterung ebenfalls unterhalb der Maskierschwelle bleibt. Gelingt dies, sind alle störenden Verzer-

rungen des Nutzsignals und der Störung unhörbar. Im Gegensatz zu den bereits in der Literatur aufgeführten psychoakustischen Verfahren wird somit eine zusätzliche Minimierung aller Signalverzerrungen erreicht. Die Bestimmung der Mithörschwelle erfolgt mittelbar durch LPC- oder Kepstralschätzung der Leistungsdichte des Nutzsignals. Die Störschätzung erfolgt mit der hier vorgestellten CA-Methode. Instrumentelle Experimente und Hörvergleiche haben belegt, daß mit dem *CA-CA* und *CA-LPC*-Verfahren die besten Reduktionsergebnisse bei bestem Höreindruck erreicht werden konnten. In umfangreichen Hörtest wurden die theoretischen Aussagen und Untersuchungen subjektiv bestätigt.

Diese Verfahren sind wegen Ihrer Eigenschaften für die Anwendung im Fahrzeug mit besonders kleinen Signal-Störabständen geeignet. Sie sind nach erster Abschätzung auf herkömmlichen digitalen Signalprozessoren in Echtzeit umsetzbar. Die Betrachtung mehrkanaliger Ansätze wurde in dieser Arbeit nicht weiter durchgeführt, da diese Systeme im Fahrzeug einen überproportionalen Aufwand und hohe Kosten verursachen.

Für die Zukunft ist ein ähnlicher psychoakustischer Ansatz für die Systemidentifikation denkbar. Üblicherwiese ist es besonders schwierig, die Übertragungsstrecken zwischen mehreren Fahrzeuglautsprechern und einem oder mehreren Mikrophonen systematisch zu schätzen. Aufgrund der im allgemeinen instationären Raumimpulsantwort auf den verschiedenen Systempfaden ergibt sich mit den bisher etablierten Verfahren (z.B. Normalized-Mean-Square-Algorithmus, Recursive-Least-Square-Algorithmus etc.) ein hoher Realisierungsaufwand. Außerdem besitzen Sprachsignale wegen ihrer Formantstruktur nur schlechte Adaptionseigenschaften bei Anwendung dieser Algorithmen.

Aus diesem Grund wurde ein Verfahren entwickelt [131], das die psychoakustischen Eigenschaften des menschlichen Gehörs nutzt, um spezielle orthogonale Sequenzen zusammen mit dem Lautsprechersignal zu senden. Dabei werden solche Sequenzen verwendet, die wegen ihrer besonderen Korrelationseigenschaften die Konvergenz der oben beschriebenen Identifikationsverfahren verbessern und vor allem eine selektive Identifikation der einzelnen Lautsprecher-Mikrophon-Übertragungsfunktionen erlauben. Die orthogonalen Sequenzen werden so gestaltet, daß sie unterhalb der Mithörschwelle $R_T(n, m)$ verbleiben und damit nicht wahrnehmbar sind. Auf die Implementation und Untersuchung dieses Verfahrens wurde in dieser Arbeit verzichtet. Die psychoakustische Methode zur selektiven Identifikation der Raumimpulsantwort eines Fahrgastraumes wurde mit [132] patentrechtlich geschützt.

A Anhang

A.1 Hörproben und Beispiele

Auf der beiliegenden CD wurden verschiedene Hörbeispiele für die Präsentation oder den Hörtest und Vergleich zusammengestellt. Wegen des recht großen Dynamik- und Frequenzbereichs der Tonbeispiele ist beim Abspielen der Hörproben auf die eingestellte Lautstärke zu achten. Es sollte möglichst kein Kopfhörer verwendet werden. So werden Lautsprecher und Gehör ausreichend geschützt. Folgende Daten sind auf der *extended* CD zu finden:

- *Dissertation als strukturiertes PDF-File mit farbigen Diagrammen und Abbildungen*: Die vorliegende Arbeit kann mit dem Acrobat Reader© geöffnet werden. Eine übersichtliche Dokumentstruktur erleichtert dem Leser das Navigieren und Suchen.
- *Testdaten im WAV-Format für den Vergleich zu anderen Verfahren*: Dazu gibt es im Verzeichnis `/Hoervergleich/` zwei Verzeichnisse mit Testdaten. `/db_5/` enthält verschiedene Tonproben mit einem Signal-Störverhältnis von 5dB. `/db_0/` enthält dieselben Daten mit einem Signal-Störverhältnis von 0dB.
- *Audiostruktur mit Hörbeispielen*: Die CD beinhaltet zudem verschiedene Hörbeispiele, die in jedem handelsüblichen CD-Player abspielbar sind. Folgende Hörbeispiele wurden zusammengestellt:
- Hörbeispiele zur Erklärung der Lautheit, Frequenzgruppe und Wahrnehmung differentieller Änderung des Schalls, aus [172]. Die jeweilige Nummer links der Beschreibung gibt die Nummer des Audiotitels auf der CD an.

    (1) *Zunahme der Lautheit von Schall gleicher Intensität mit anwachsender Bandbreite*: Neun mal werden abwechselnd ein Sinuston und ein bandbegrenztes Rauschen dargeboten. Der Sinuston hat die Frequenz 1000 Hz. Die Mittenfrequenz des Rauschens beträgt ebenfalls 1000 Hz. Dabei wächst die Bandbreite des Rauschens von 100 Hz bis 1900 Hz an. Die Gesamtintensität bleibt konstant. Ab der dritten Darbietung ist die Bandbreite des Rauschens größer als die Frequenzgruppe. Damit gelangen Anteile des Rauschens in benachbarte Frequenzgruppen und bewirken auch dort einen Beitrag zur Lautheitsempfindung. Die empfundene Lautstärke des Rauschen sollte daher ab dem dritten Beispiel höher sein als die des Sinustones.

(2) *wahrnehmbarer Frequenzunterschied*: Zehn Gruppen von zwei Tonpaaren werden dargeboten. Bei jedem Tonpaar ist der zweite Ton entweder höher oder tiefer als der erste Ton. Die Reihenfolge der Tonpaare variiert zufällig. Der Frequenzunterschied zum zweiten Ton wird von 9 Hz bis 0 Hz schrittweise verringert. Durch Abzählen der Gruppen, bei denen ein Frequenzunterschied gerade noch wahrgenommen wird, kann die individuelle Schwelle für Frequenzunterschiede bestimmt werden.

- Psychoakustische Verdeckungseffekte an Beispielen, aus [172]. Hierbei werden noch einmal die in Abschnitt 3.3 behandelten psychoakustischen Maskiereffekte an Hörproben deutlich gemacht.

(3) *Mithörschwelle eines Sinustones, verdeckt durch weißes Rauschen*: In der Demonstration wird ein 2 kHz-Sinuston durch weißes Rauschen verdeckt. Hierbei wird zunächst der Sinuston abgespielt. Anschließend wird der schrittweise um jeweils 3dB gedämpfte Sinuston durch weißes Rauschen verdeckt.

(4) *Mithörschwelle eines Sinuston, durch verschiedene bandbegrenzte Rauschsignale*: Die gleiche Sequenz wird jetzt mit bandbegrenzten Rauschsignalen mit der Bandbreite 7000 Hz und 1000 Hz überlagert. Die Bandbreiten sind größer als die Frequenzgruppe in der sich der Sinuston befindet. Beim jeweils zweiten Rauschsignal wird der Sinuston deutlich stärker verdeckt, da sich hier Rauschanteile stärker um den Sinuston konzentrieren als beim ersten Rauschsignal.

(5) *Maskierung komplexer Töne*: Die vorgespielte Melodie enthält jeweils den Sinuston und die überlagerte vierte, fünfte und sechste Harmonische. Nun wird ein Hochpaßrauschen überlagert. Dies macht die Oberwellen unhörbar, nur die Sinustöne bleiben weiter hörbar.

- Einfluß der Parameter auf die vorgestellten Verfahren zur Geräuschreduktion. Dabei wird als Störgeräusch *rosa Rauschen* mit $SNR = 0\text{dB}$ verwendet. In den einzelnen Hörproben wird jeweils der Spectral Floor b geändert. Demnach ändert sich auch die Adaption des Oversubtraction-Faktors a und die psychoakustische Verdeckung der neu entwickelten Filterregel:

(6) *CA-CA-Verfahren* mit adaptiver Einstellung des Oversubtraction-Faktors a je nach Vorgabe des spectral floors: $b = 0, b = 0.1, b = 0.2$

(7) *CA-LPC-Verfahren* mit adaptiver Einstellung des Oversubtraction-Faktors a je nach Vorgabe des spectral floors: $b = 0, b = 0.1, b = 0.2$

- Vergleich der verschiedenen Verfahren bei unterschiedlichen Geräuscharten und Signal-Stör-Verhältnissen (SNR). Deutlich sind in beiden Audiodemonstrationen die unterschiedlichen Verfahren, mit Reihenfolge lt. Tabelle 6.2, zu erkennen.

(8) Mischung des Sprachsignals mit den verschiedenen Geräuschen: $SNR = 5\text{dB}$, rosa Rauschen, konstante Fahrt, dynamische Fahrt, Regenfahrt.

(9) Mischung des Sprachsignals mit den verschiedenen Geräuschen: $SNR = 0\text{dB}$, rosa Rauschen, konstante Fahrt, dynamische Fahrt, Regenfahrt.

Quellen- und Literaturverzeichnis

[1] Ades, H. W.; Engström, H.: Anatomy of the inner ear, Handbook of Sensory Physiology. Springer-Verlag Berlin-Heidelberg, 1974

[2] Ahmed, M.: Comparision of Noisy Speech Enhancement Algorithms in Terms of LPC Perturbation. IEEE Trans. on ASSP, Vol. 37 (1989), No. 1, pp 121-125

[3] Ahmed, S.A.; Cruz, J.R.: Complex System Identifikation Methods for fast Echo Canceller Initialization. Proc. of ICASSP ’92, Vol. IV, pp. 525-528

[4] Ali, M.: Stereophonic Acoustic Echo Cancellation using time-varying All-Pass Filtering for Signal Decorrelation. ICASSP 98, Vol 6 CD-ROM

[5] Allen, J.; Berkley, D.; Blauert, J.: Multimicrophone Signal-Processing Technique to remove Room Reverberation from Speech Signals. J. Acoust. Soc. Am., Vol. 62 (1977), pp. 912-915

[6] Alrutz, H.: Über die Anwendung von Pseudorauschfolgen zur Messung an linearen Übertragungssystemen. Dissertation Universität Göttingen, 1983

[7] Alrutz, H.; Gottlob, D.: Bestimmung raumakustischer Impulsantworten mit Pseudo-Rauschsignalen. Fortschritte der Akustik, DAGA-76, S. 267-270

[8] Amand, F.; Benesty, J.; Gilloire, A.; Grenier, Y.: A Fast Two Channel Projection Algorithm for Stereophonic Acoustic Echo Cancellation. IEEE Proc. ICASSP ’96, pp. 949-952

[9] Antweiler, C.: Adaption of Acoustic Echo Cancellers exploiting Spectral characteristics of room impulse responses. IEEE Proc. IWAENC ’97, pp. 57-60

[10] Antweiler, C.: Orthogonalisierende Algorithmen für die digitale Kompensation akustischer Echos. Verlag der Augustinus-Buchhandlung, 1995.

[11] Antweiler, C.: Simulation of Time Variant Room Impulse Responses. IEEE Proc. ICASSP 97, pp. 3031- 3034

[12] Antweiler, C.; Dörbecker, M.: Perfect Sequence Exitation of the NLMS Algorithm and its Application to Acoustic Echo Control. Annales des Télécommunication, Vol. 49 (1994), No.7/8, pp. 386-397

[13] Ayad, B.; LeBouquin-Jeannés, R.: Acoustic Echo and Noise Reduction: A novel Approach. IEEE Proc. IWAENC ’97, pp. 168-171

[14] Azirani, A.A.; LeBouquin-Jeannés, R.; Fauson, G.: Optimizing Speech Enhancement by exploiting masking Properties of the Human Ear. IEEE Proc. ICASSP ’95, pp. 800-803

[15] Baumann, U.: Ein Verfahren zur Erkennung und Trennung multipler akustischer Objekte. H. Utz Verlag Wissenschaft München, 1995

[16] Bappert, V., Blauert, J.: Auditory quality evaluation of speech-coding systems. Acta Acustica, No. 2, S.49-58, 1994.

[17] Becker, D.: Untersuchungen zur Störreduktion bei der automatischen sprecherabhängigen Erkennung von isoliert gesprochenen Einzelwörtern. Dissertation TU-Berlin, 1992

[18] Becker, T.; Hänsler, E.; Schultheiß, U.: Probleme bei der Kompensation akustischer Echos. Frequenz 36 (1980), Band 6, S. 142-148

[19] Bell, K.L., Ephraim, Y., Trees, H.L.: Robust Adaptive Beamforming using Data dependent Constraints. ICASSP 98, CD-ROM

[20] Beranek, L.L.: Acoustic Measurements. Wiley, New York, 1949

[21] Berouti, M.; Schwartz, R.; Makhoul, J.: Enhancement of Speech corrupted by Acoustic Noise. Proceedings of ICASSP, 1979, pp. 208-212

[22] v. Bismarck, G.: Sharpness as an attribute of the timbre of steady sounds. Acustica 1974, S. 146-159

[23] Bjarnason, E.: Active Noise Cancellation using a modified Form of the Filtered-X LMS Algorithm. Signal Prosessing IV, EUSIPCO-92, Vol. 2, (1992), S. 1053-1056

[24] Blauert, J.: Räumliches Hören. Hirzel, Stuttgart, 1974

[25] Bodden, M., Rateitschek, K.: Noise-robust speech recognition based on a binaural auditory model. Proc. of WABSP 1996, pp. 291-296

[26] Boll, S.: Surpression of Acoustic Noise in Speech using Spectral Substraction. IEEE Trans. on ASSP, Vol 27 (1979), pp. 113-120

[27] Borish, J.: Self-Constrained Crosscorrelation Program for Maximum-Length Sequences. J. Audio Eng. Soc., Vol. 33 (1985), No. 11

[28] Borish, J.; Angell, J.B.: An Efficient Algorithm for Measuring the Impulse Response using Pseudorandom Noise. J. Audio Eng. Soc., Vol. 31 (1983) pp. 478-488

[29] Bronzel, M.: Aktive Beeinflussung nicht stationärer Schallfelder mit adaptiven Digitalfiltern. Dissertation Universität Göttingen, 1993

[30] Burkhardt, T.; Strube, H.: Adaptive Verfahren zur Entstörung von Sprache im PKW. DAGA ’90, Fortschritte der Akustik, 1990, S. 1135-1138

[31] Carter, G.: Coherence and Time Delay Estimation. Proc. IEEE, Vol. 75 (1987), No. 2, pp. 236-255

[32] Cattermole, K.W.: Statistische Analyse und Struktur von Information. VCH, 1988.

[33] Chai, Y.: Modellbasierte Störreduktion bei Sprachsignalen. Fortschr. Berichte VDI-Reihe 10 Nr. 297, Düsseldorf 1994.

[34] Chan, C.F., Hiu, W.K.: Quality Enhancement of Narrowband CELP-Coded Speech via Wideband Harmonic Re-Synthesis. ICASSP 97, pp. 1187-1190

[35] Chan, D.C.B.; Rayner, P.J.W.; Godstill, S.J.: Multi-Channel Signal Separation. IEEE Proc. ICASSP ’96, pp. 649-652

[36] Cheng, Y.; O’Shaughnessy, D.: Speech Enhancement Based Conceptually on Auditory Evidence. IEEE Trans. on SP, Vol. 39 (1991), No. 9, pp. 1943-1954

[37] Cioffi, J.M.; Kailath, T.: Fast, Recursive-Least-Squares Transversal Filters for Adaptive Filtering. IEEE Trans. on ASSP, Vol. 32; (1984), No. 2, pp. 304-337

[38] Cohe, M.A.: Time-Frequency-Analysis, Tutorial TFTS 1998, Pittsburg

[39] Cohen, M.A.; Grossberg, S.; Wyse, L.L.: A spectral network model of pitch perception. J. Acoust. Soc. Am. Band 98 (1995), pp. 862-879

[40] Cohen, L., Nelson, D., Umesh, S.: Fitting the Mel Scale. ICASSP 99 Vol.2 CD-ROM

[41] Cohn, M.; Lempel, A.: On fast M-sequence transforms. IEEE Trans. on Inf. Theory, Vol. 23 (1977), pp. 135-137

[42] Cooke, R.K.: Modelling Auditory Processing and Organisation. Cambridge, 1993

[43] Deisher, M., Spanias, A.S.: HMM-Based Speech Enhancement using Harmonic Modelling. Proc. ICASSP-97, Munich. pp. 1381-1384

[44] Deller, J.R.; Proakis, G.; Hansen, H.L.H.: Discrete-Time Processing of Speech Signals. Prentice-Hall, 1993

[45] Drews, M.: Mikrofonarrays und mehrkanalige Signalverarbeitung zur Verbesserung gestörter Sprache. Dissertation. Berlin 1999

[46] Edelmann, G.M.; Hall, W.E.; Cowan, W.M.: Auditory Function. Wiley, New York, 1988

[47] Egan, J.P.; hake, H.W.: On the masking pattern of a simple auditory stimulus. J. Acoust. Soc. Am. Band 22 (1950), pp. 622-630

[48] Ephraim, Y.: Statistical-Model-Based Speech Enhancement Systems. Proc. IEEE, Vol. 80 (1992), No. 10, pp. 1526-1555

[49] Ephraim, Y.; Malah, D.: Speech Enhancement using a Minimum Mean-Square Error Short-time Spectral Amplitude Estimator. IEEE Trans. on ASSP, Vol. 32 (1984), No. 6, pp. 1109-1121

[50] Ephraim, Y., Malah, D., Juang, B.H.: On the application of hidden markov models for enhancing noise speech. IEEE Trans. on Acoust. Speech and Signbal Proc. vol. 37, pp. 1846-1856 Dec. 1989

[51] ETSI Rec. GSM 06.92: Voice-Activity-Detector, Valbourne 1989.

[52] Fant, G.: Acoustic Theory of Speech Production. The Hague, 1960

[53] Fischer, S., Kammeyer, K.D.: Broadband Beamforming with Adaptive Postfiltering for Speech Acquisition in noisy Environments. ICASSP 99 CD-Rom

[54] Fischer, S.: Adaptive Mehrkanalgeräuschunterdrückung bei gestörten Sprachsignalen unter Berücksichtigung der räumlichen Kohärenz des Geräuschfeldes. Diss. Universität Bremen, 1996. Verlag Shaker, Aachen

[55] Fliege, N.: Multiraten-Signalverarbeitung: Theorie und Anwendungen. Teubner, 1993.

[56] Fraunhofer Gesellschaft zur Förderung der angewandten Forschung: Digital Encoding Process. United States Patent No. 5579430, 1996.

[57] Frenzel, R.: Freisprechen in gestörter Umgebung. VDI-Verlag, 1992. Fortschritt-Berichte Reihe 10 Nr. 228

[58] Frenzel, R.; Hennecke, M.E.: Using Prewhitening and Stepsize Control to improve the performance of the LMS Algorithm for Acoustic Echo Compensation. Proc. ISCAS 1992, pp. 1930-1932

[59] Furui, S.: Digital Speech Processing, Synthesis, and Recognition. New York, M. Dekker, 1989.

[60] Gannot, S.; Burshtein, D.; Weinstein, E.: Iterative and Sequential Kalman Filter based Speech Enhancement Algorithms. Trans. on SAAP, Vol. 6, 1998, No. 4, pp. 373-385

[61] Gay, S.L.: Dynamically Regulized Fast RLS with Application to Echo Cancellation. IEEE Proc. ICASSP ’96, pp. 957-960

[62] Gierl, S.: Geräuschreduktion bei Sprachübertragung mit Hilfe von Mikrofonarraysystemen. Dissertation Universität Karlsruhe, 1990

[63] Gierlich, H.W.: A measurement technique to determine the transfer characteristics of hands-free telephones. Signal Processing 27, 1992, pp. 281-300

[64] Gierlich, H.W., Kettler, F., Krebber, W., Dietrich, E.: Quality evaluation procedures for hand-free telephones. Proc. ITG-EURASIP Workshop Quality Assessment Speech, Audio and Image Communication, Darmstadt, S.30/31, 1996.

[65] Godsill, S.J.; Rayner, P.J.W.: Robust Noise Reduction for Speech and Audio Signals. IEEE Proc. ICASSP ’96, pp. 625-628

[66] Goh, Z.; Tan; K.C.; Tan, B.T.G.: Postprocessing Method for Surpressing Music Noise Generated by Spectral Substraction. IEEE Trans. on SAAP, Vol. 6, 1998, No. 3, pp. 287-292

[67] Green, D.M.: Additivity of Masking. J. of Acoust. Soc. Am., Nr. 41 Vol.6, Jan 1967

[68] Greenberg, J.E.: Modified LMS Algorithms for Speech Processing with an Adaptive Noise Canceller. IEEE Trans. on SAAP, Vol. 6, 1998, No. 4, pp. 338-351

[69] Guelou, Y.; Benemar, A.; Scalart, P.: Analysis of two Structures for combined Acoustic Echo Cancellation and Noise Reduction. IEEE Proc. ICASSP ’96, pp. 637-640

[70] Gustafsson, S., Martin, R.: Combined Echo Control and Noise Reduction for Mobile Communications. Proc. Eurospeech 97, pp. 1403-1406

[71] Gustafsson, S., Jax, P., Kamphausen, A., Vary, P.: A Postfiler for Echo and Noise Reduction avoiding the Problem of Musical Tones. Proc. ICASSP 99, Vol. 2, pp. n/a

[72] Gustafsson, P., Jax, P., Vary, P.: A Novel Psychoacoustically motivated Audio Enhancement Algorithm preserving Background Noise Characteristics. ICASSP 98 Vol 2, CD-ROM

[73] Haykin, S.: Adaptive Filter Theory. Prentice-Hall, 1996

[74] Heinbach, W.: Datenreduktion von Sprache unter Berücksichtigung von Gehöreigenschaften. ntz, Band 9 (1987), S. 327-333

[75] Hirsh, I.J.; Ward, W.D.: Recovery of the auditory treshold after strong acoustic stimulation. J. Acoust. Soc. Am., Band 24 (1952), pp. 131-141

[76] Hoffmann, J.: Matlab und Simulink. Addison Wesley, 1998

[77] Hoshuyama, O., Sugiyama, A., Hirano, A.: A Robust Adaptive Microphone Array with improved spatial Selectivity and its Evaluation in a real Environment. ICASSP 1997, pp. 367-370

[78] Huhn, J.; Jentschel, H.J.: Kombination von Geräuschreduktion und Echokompensation beim Freisprechen. Nachrichtentechnik, Elektronik, Vol. 43 (1993), No. 6, pp. 274-280

[79] Hänsler, E.: Adaptive Echo Compensation applied to the Hands-Free Telephone Problem. Proc. ISCAS 1990, pp. 279-282

[80] Hänsler, E.: Statistische Signale: Grundlagen und Anwendungen. Springer-Verlag, Berlin-Heidelberg-New York, 1997

[81] Hänsler, E.: The hands-free telephone problem - An annotated bibliography. Signale Processing 27, 1992, pp. 259-271

[82] Humes, L.E., Jesteadt, W.: Models of the additivity of masking. J. Acoust. Soc. Am. Vol. 85 Nr. 3, pp. 1285-294, 1989.

[83] Humes, L.E., Lee, L.W.: Two experiments on the spectral boundary conditions for nonlinear additivity of simultaneous masking. J. Acoust. Soc. Am. Vol. 92 Nr. 5, pp. 2598-2606, 1992.

[84] Humes, L.E., Cokely, C.G.: Two experiments on the temporal boundaries for the nonlinear additivity of masking. J. Acoust. Soc. Am., Vol. 94, Nr. 5, pp. 2553-2559, 1993.

[85] Ipatov, V.P.: Contribution to the Theory of Sequences with perfect Autocorrelation Properties. Radio Eng. Elect. Phys. 25, 1980, pp. 31-34

[86] Ipatov, V.P.: Ternary Sequences with Ideal Periodic Autocorrelation Properties. Radio Eng. Elect. Phys. 24, 1979, pp. 75-79

[87] ISO/IEC: Coding of moving Pictures and Associated Audio, Psychoacoustic Models for Lower Sampling Frequencies. Publ. Nr. ISO/IEC13818-3:1994(E) Part 3 Annex D, 1994.

[88] ITU-T Rec. P.810: Modulated noise reference unit. Genf, Februar 1996

[89] Johnson, J.D.: Transform Coding of Audio Signals Using Perceptual Noise Criteria. IEEE Journal on Selected Areas of Communcation, Vol 6, No. 2, pp. 314-323, 1998.

[90] Kajita, S.; Takeda, K.; Itakura, F.: A Binaural Speech processing method using Subband-Crosscorrelation Analysis for noise robust Recognition. IEEE Proc. ICASSP ’97, pp. 1243-1247

[91] Kates, J.M.: Speech enhancement based on a sinusoidal model. J. Speech Hear. Res., Band 27 (1994), pp. 449-464

[92] Katsaggelos, K.A.: Image and Video recovery and Enhancement Techniques. IASTED Tutorial SIP’98, Las Vegas, 1998

[93] Kellermann, W.: A Self Steering Digital Microphone Array. Proc. of ICASSP 91, pp. 3581-3584

[94] Kobler, J.: Minimum-Varianz-Schätzer zur Geräusch-Reduktion bei der Einzelworterkennung. Dissertation Universität Karlsruhe, 1994

[95] Krahé, D.: Ein Verfahren zur Datenreduktion bei digitalen Audiosignalen unter Ausnutzung psychoakustischer Phänomene. Rundfunktechnik, Band 30 (1986), S. 117-123

[96] Krolikowski, R., Czyzewski, A.: Noise Reduction in Acoustic Signals Using Perceptual Coding. 137th Meeting Acoust. Soc. of Am., Berlin, Germany. 1997

[97] Kroschel, K.: Statistische Nachrichtentheorie (2. Teil). Springer-Verlag, Berlin-Heidelberg-New York, 1974

[98] Kroschel, K.: Umgebungsgeräuschreduktion bei Sprachkommunikationssystemen. Frequenz 42 (1988), S. 79-84

[99] Kwong, R.H., Johnston, E.W.: A viariable step size LMS algorithm. IEEE Trans. on SP, Vol. 40 (1992), No. 7, pp. 1633-1642

[100] Lang, M., Guo, H., Odegard J.E., Burrus, C.S., Wells, R.O.: Noise reduction using an undecimated discrete wavelet. IEEE Signal Processing letters, Vol.3, 1996. pp. 10-12

[101] Lang, M., Guo, H., Odegard J.E., Burrus, C.S., Wells, R.O.: Nonlinear processing of shift invariant DWT for noise reduction. Proc. of SPIE 1995, pp. 640-651

[102] Lee, K.Y., Shirai, K.: Recursive estimation for speech enhacement using the hidden filer model. Proc. Acoust. Soc. Japan 1995, pp. 63-64

[103] Lim, J.S.; Oppenheim, A.V.: Enhancement and Bandwidth Compression of Noisy Speech. Proceedings of the IEEE, Vol. 67, No. 12 Dec. 1979, pp. 1586-1604

[104] Linhard, K.: Adaptive Geräuschreduktion im Frequenzbereich bei Sprachübertragung. Dissertation, Universität Karlsruhe, 1988

[105] Linhard, K., Haulick, T.: Nichlineare Glättung und Geräuschreduktion bei gestörten Sprachsignalen. Proc. Aachener Kolloquium Signaltheorie, 1997. pp. 251-254

[106] Lüke, H.D.: Korrelationssignale. Springer-Verlag, Berlin-Heidelberg-New York, 1992

[107] Lüke, H.D.: Sequences and Arrays with perfect periodic Correlation. IEEE Trans. on Aerospace and Electronic Systems, Vol. 24, No. 3, 1984, pp. 287-294

[108] Lüke, H.D.: Ungerade perfekte Binärfolgen für die Kanalmeßtechnik. 8. Aachener Kolloquium Signaltheorie, 1994, S. 111-114

[109] Lutfi, R.A.: Additivity of simultaneous masking. J. Acoust. Soc. of Am. No. 73, pp. 262-267, 1983.

[110] Lutfi, R.A.: A Power-Law Transformation Predicting Masking by Sounds with Complex Spectra. J. Acoust. Soc. Am. No. 77 Vol. 6 June 1985

[111] Markel, J.; Gray, A.: Linear Prediction of Speech. Springer-Verlag, New York, 1976

[112] Marro, C.; Mahieux, Y.; Simmer, K.U.: Analysis of Noise Reduction and Dereverberation Techniques based on Microphone Arrays with Postfiltering. IEEE Trans. on SAAP, Vol. 6, 1998, No. 3, pp. 240-259

[113] Martin, R.: Freisprecheinrichtungen mit mehrkanaliger Echokompensation und Störgeräuschreduktion. Verlag Augustinus Buchhandlung, 1995.

[114] Martin, R.: Spectral Substraction Based on Minimum Statistics. Proc. EUSIPCO-94, Edinburgh, 1994. pp. 1182-1185

[115] Martin, R.: An Efficient Algorithm to Estimate the Instantaneous SNR of Speech Signals. Proc. EuroSpeech 1993, pp. 1093-1096

[116] Matrouf, D., Gauvain, J.-L.: Using AR HMM State-Dependent Filtering for Speech Enhancement. ICASSP 98, Vol 2. CD-ROM

[117] Matt, H.J., Walker, M.: Acoustic Echo and Line Echo Supression using a Combination of a Shorzt Length Adaptive Filter and a Compander. Proc. Aachener Kolloguium Signaltheorie, 1997. pp. 271-274

[118] McCaslin, S.: Echo Cancellation for Speech Applications. 3rd Int. Workshop on Acoustic Echo Control, Plestin les Greves, 1993, pp. 203-211

[119] McOlash, S. M., Niederjohn, R. J., Heinen, J. A.: A spectral subtraction method for enhancement of speech corrupted by nonwhite, nonstationary noise. IEEE Proc. 21. International Conference on Industrial Electrical Control and Instrumentation, 1976, pp. 872-877

[120] Meyer, J.; Simmer, K.U.; Kammayer K.D.: Comparison of one- and two-channel noise estimation techniques. IEEE Proc. IWAENC ’97, pp. 17-20

[121] Meyer, J., Sydow, C.: Noise Cancelling for Microphone Arrays, ICASSP 1997, pp. 211-214

[122] Meyer, J., Simmer, U.: Multi-Channel Speech Enhancement in a Car Environment using Wiener-Filtering and Spectral Substraction. ICASSP 98 CD-Rom

[123] Moore, B.C.J.: An Introduction to the Psychology of Hearing. Academic, London, 1992

[124] Naylor, P.A.; Tanrikulu, O.; Constantinides, A.G.: A Better Understanding and an Improved Solution to the Specific Problems of Stereophonic Acoustic Echo Cancellation. IEEE Trans. on SAAP, Vol. 6, 1998, No. 2, pp. 156-165

[125] Ohm, D.: Untersuchungen zur nichtlinearen Prädiktion des Innraumgeräusches von Kraftfahrzeugen durch Neuronale Netze

[126] Oppenheim, A.V.; Willsky, A.S.: Signals and Systems. Prentice-Hall, NJ, 1983

[127] Orgren, A.C.; Dasgupta, S.; Rohrs, C.E.; Malik, N.R.: Noise Cancellation with improved Residuals. IEEE Trans. on SP, Vol. 39, 1991, No. 12, pp. 2629-2639

[128] Parker, S.P.: Circuits, Systems and Signal Processing. Birkhäuser Boston 1998

[129] Pandit, M.: Grundlagen und Anwendungen der Theorie stochastischer Prozesse. Universität Kaiserslautern. Umdurck zur Vorlesung.

[130] Peters, M.: Binaural Bark Subband Preprocessing of Nonstationary Signals for Noise Robust Speech Feature Extraction. Signal and Image Processing, International Conference on Acoustics, Speech, and Signal Processing, Phoenix, AZ, 1999. Vol. 1 Paper No. 1874

[131] Peters, M.: Psychoacoustical Excitation of the (N)LMS Algorithm for Acoustical System Identification. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, 1999. Proceedings pp. 211-214

[132] Peters, M.: Verfahren und Vorrichtung zur Ermittlung der akustischen Raumeigenschaften insbesondere eines Fahrgastraumes. Patentschrift BMW AG, Forschungs- und Ingenieurzentrum, München 1999. Pat. DE 19933317.3

[133] Peters, M., Weinmann, U.: Verfahren zur drahtlosen Übertragung von Nachrichten zwischen einem fahrzeuginternen Kommunikationssystem und einem fahrzeugexternen Zentralrechner. Patentschrift BMW AG, Forschungs- und Ingenieurzentrum, München 1999. Pat. DE 19933318.1

[134] Peters, M., Weinmann, U.: Spracherkennungssystem und Verfahren zur Spracherkennung vorgegebener Sprachmuster, insbesondere Sprachsteuerung. Patentschrift BMW AG, Forschungs- und Ingenieurzentrum, München 1999. Pat. DE 19933323.8

[135] Peters, M.: Bedienkonzepte im Fahrzeug - Das adaptive und intuitive HMI. Verein Deutscher Ingenieure, Fachtagung „Elektronik im Kraftfahrzeug“, Baden-Baden, 2001, to be published.

[136] Peters, M., Herrler, M., Zeller, M., Spreng, M.: iDrive - Ein neuer Maßstab in der Fahrzeugbedienung. ATZ 10/2001, to be published.

[137] Peters, M., Spreng, M: iDrive - Das neue Bedienkonzept. BMW AG München. Fachtagung Automobilelektronik „Fortschritt und Zukunft in der Automobil Elektronik“, Stuttgart, 06/2001, to be published.

[138] Peters, M.: Nichthörbare Korrelationssignale für die Schätzung akustischer Raumimpulsantworten und die Verbesserung des Adaptionsverhaltens des NLMS-Algorithmus. Universität Kaiserslautern, Nachrichtentechnisches Seminar 02/99

[139] Peters, M.: Noise robust speech recognition and speech enhancement in car environment. Universität Kaiserslautern, Nachrichtentechnisches Seminar 03/98

[140] Peters, M.: Die Realisierung einer mittelratigen störunterdrückenden Funkübertragung im ISM-Band. Universität Kaiserslautern, Lehrstuhl für Nachrichtentechnik Diplomarbeit, 1996

[141] Peters, M.: Space Vektor Control for Asynchronous Machines and Digital Signal Processing on AT&T’s DSP32c. Michigan State University, Dept. Electrical Engeneering, 1993. Junior Thesis.

[142] Peters, R.W.; Hall, J.W.: Change in the pitch of a complex tone following its association with a second complex tone. J. Acoust. Soc. Am. Band 71 (1982), pp. 142-146

[143] Peters, R.W.; Hall, J.W.: Pitch for nonsimultanous successive harmonics in quiet and noise. J. Acoust. Soc. Am. Band 69 (1981), pp. 509-513

[144] Peters, R.W.; Moore, B.C.J.; Glasberg, B.R.: Pitch of components of complex tones. J. Acoust. Soc. Am., Band 24 (1952), 175-184

[145] Pierre, D.A.: Optimization Theory with Apllications, 1969, Wiley, pp. 561-562

[146] Preuss, R.D.: A Frequency Domain Noise Cancelling Preprocessor for Narrowband Speech Communication Systems. Poceedings of ICASSP, 1979, pp. 212-215

[147] Rabenstein, R.: Filterstrukturen für zeitvariante rekursive Systeme. Tagungsband 5. Aachener Kolloqium, Sept. 1984, S. 186-189

[148] Rabiner, L.R.; Schaefer, R.W.: Digital Signal Processing of Speech Signals. Prentice-Hall, Englewood Cliffs, NJ

[149] Raman, V.R.; Cromack, M.R.: Fast echo cancellation in a voice-processing system. Proc. ICASSP ’92, Vol. 4, pp. 513-516

[150] Reich, W.: Adaptive Systeme zur Reduktion von Umgebungsgeräuschen bei Sprachübertragung. VDI-Verlag, 1986. Fortschritt-Berichte, Reihe 10 Nr. 63

[151] Reininger, H.; Kuhn, C.: Signalverbesserung durch gehörgerechte Spektrale Subtraktion. Proc. 9. Aachener Kolloquium Signaltheorie, 1997, S. 259-262

[152] Rife, D.; Vanderkooy, J.: Transfer-Function Measurement with Maximum-Length Sequences. J. Audio Eng. Soc., Vol. 37, 1989, No. 6, pp. 419-444

[153] Robinson, E.A.; Silvia, M.T.: Digital Signal Processing and Time Series Analysis. Holden-Day, 1978, pp. 363-366

[154] Rosenblith, W.A.; v. Bésésky, G.: The early history of Hearing-Observations and Theories. J. Acoust. Soc. Am 74 (1983), pp. 428-432

[155] Rupp, M.: Über die Analyse von Gradientenverfahren zur Echokompensation. VDI-Verlag, 1993. Reihe 10 Nr. 242

[156] Rühl, H.; Dobler, S.; Weith, J.; Meyer, P.; Noll, A.; Hamer, H.; Piotrowski, H.: Speech recognition in Noisy Car Environment. Speech Communication, Vol. 10 (1991), No. 1

[157] Ryan, J.G., Goubran, R.A.: Near-Field Beamforming for Microphone Arrays. ICASSP 98 CD-Rom

[158] Scalart, P.; Filho, J.V.: Speech Enhancement based on apriori Signal to Noise Estimation. IEEE Proc. ICASSP ’96, pp. 629-632

[159] Schirmacher, R.: Schnelle Alorithmen für adaptive IIR-Filter und ihre Anwendung in der aktiven Schallbeeinflussung. Dissertation Universität Göttingen, 1995

[160] Schlang, M.F.: Methoden zur Störschallunterdrückung bei ortsungebundener Spracheingabe in Räumen. Disseration TU München, 1991.

[161] Shimauchi, S.; Makino, S.: Stereo Echo Cancellation using imaginary input-output relationships. IEEE Proc. ICASSP ’96, pp. 941-944

[162] Sondhi, M.M.; Schmidt, C.E.; Rabiner, L.R.: Improving the Quality of a Noisy Signal. BSTJ, Vol. 60, Oct. 1981, pp. 1847-1853

[163] Stearns, S.D.: Digitale Verarbeitung analoger Signale. Oldenbourg-Verlag, 1991

[164] Sörqvist, P., Händel, P., Ottersen, B.: Kalman Filtering for low Distortion Speech Enhancement in Mobile Communication. ICASSP 98 CD-Rom

[165] Steinbuch, K.; Rupprecht, W.: Nachrichtentechnik. Springer-Verlag, Berlin-Heidelberg-New York, 1992

[166] Stevens, S.S.: Psychophysics. Wiley, New York, 1975

[167] Strang, G.; Nguyen, T.: Wavelets and Filter Banks. Wellesley-Cambridge-Press, 1996

[168] Strobe, B.; Alwan, A.: A Model of dynamic auditory Perception and its Application to robust Speech Recognition. IEEE Proc. ICASSP ’96, pp. 37-40

[169] Stumpf, C.: Tonpsychologie. Hirzel, Leipzig, 1890

[170] Sydow, C.: Selfsteering Microphone Array Systems. Dissertation Darmstadt 1996, VDI Forschungsberichte, Reihe 10, Nr. 429

[171] Tarrab, M.; Feuer, A.: Convergence and Performance Analysis of the NLMS Algorithm with Uncorrelated Gaussian Data. IEEE Trans. on Inf. Theory, Vol. 34 (1988), No. 4, pp. 680-691

[172] Terhardt, E.: Akustische Kommunikation: Grundlagen mit Hörbeispielen. Springer-Verlag, Berlin-Heidelberg-New York, 1998

[173] Thiede, T., Treurniet, W. C., Botto, R., Sporer, T., Brandenburg, K., Schmidmer, C., Keyhl, M., Beerends, J. G., Colomes, C., Stoll, G., Feiten, B.: PEAQ- der künftige ITU-Standard zur objektiven Messung der wahrgenommenen Audioqualität. Berichte der Arbeitsgruppe ITU-R TG 10/4.

[174] Tohyama, M.: Room Transfer Functions and Sound Field Control. Proc. Active ’95, pp. 15-20

[175] Tsoukalas, D.; Paraskevas, M.; Mourjopoulos, J.: Speech Enhancement using Psychoacoustic Criteria. Proc. ICASSP 93, Vol. 2, pp. 359-362

[176] Tuffy, M.A., Laurenson, D.I.: Estimating Clean Speech Thresholds for Perceptual Based Speech Enhancement. Proc. of IEEE WASPAA 1999. pp.127-130.

[177] Turbin, V.; Gilloire A.; Scalart, P.; Beaugeant, C.: Using psychoacoustic Criteria in acoustic Echo Cancellation Algorithms. IEEE Proc. IWAENC ’97, pp. 53-56

[178] Ungeheuer. G.: Elemente einer akustischen Theorie der Vokalartikulation. Springer Verlag, Berlin, 1962.

[179] v. Békésy, G.: Experiments in Hearing. McGraw-Hill, New York 1960

[180] v. Békésy, G.: Über akustische Rauhigkeit. Z. Tech. Phys. Band 16 (1935), S. 272-282

[181] v. Zitzewitz, A.: Annäherung an das ideale Freisprechtelefon mittels adaptiver Nachbildung der Übertragungsstrecke Lautsprecher-Raum-Mikrophon. Dissertation Ruhr-Universität Bochum, 1989.

[182] Vary, P.: Verfahren zur digitalen Verbesserung gestörter Sprache. TEKADE Techn. Mitteilungen 1983, S. 70-76

[183] Vary,P.: Digitale Sprachsignalverarbeitung. Teubner Stuttgart, 1998.

[184] Vaseghi, S.V.: Advanced Signal Processing and Digital Noise Reduction. Teubner Stuttgart-Leipzig-New York, 1997

[185] Vergin, R.; O’Shaughnessy, D.; Gupta, V.: Compensated Mel Frequency Cepstrum Coefficients. IEEE Proc. ICASSP ’96, pp. 323-326

[186] Virag, N.: Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System. IEEE Trans. on SAAP, Vol. 7, No. 2, 1999. pp. 126-137

[187] Wehrmann, R.; v.d.List, L.; Meissner, P.: A Noise-Insensitive Compromise Gradient Method for the Adjustment of Adaptive Echo Cancellers. IEEE Trans. on Communications, Vol. 28, No. 5 May 1980, pp. 753-759

[188] Weinstein, E.; Feder, M.; Oppenheim, A.: Multi-Channel Signal Separation by Decorrelation. IEEE Trans. on ASSP, Vol. 1 (1993), No. 4, pp. 405-413

[189] Weinstein, E.; Feder, M.; Oppenheim, A.; Buck, J.: Iterative and Sequential Algorithms for Multisensor Signal Enhancement. IEEE Trans. on SP, Vol. 42 (1994), No. 4, pp. 846-859

[190] Weinstein, E.; Feuer, A.: Convergence Analysis of LMS Filters with incorrelated Gaussian Data. IEEE Trans. on ASSP, Vol. 33 (1985), No. 1, pp. 222-228

[191] Widrow, B.: Adaptive Filters in Aspects of Network and System Theory, ed. by Kalman, R.E. and DeClaris, N. Holt, Rinehart and Winston, Inc., New York, 1971, pp. 563-587

[192] Widrow, B.; Ferrera, E.: Multichannel Adaptive Filtering for Signal Enhancement. IEEE Trans. on ASSP, Vol. 29 (1981), No. 3, pp. 766-770

[193] Widrow, B.; Glover, J.R.; McCool, J.M.; Kaunitz, J.; Williams, C.S.; Hearn, R.H.; Zeidler, J.S.; Dong, E.; Goodlin, R.C.: Adaptive Noise Cancelling; Principles and Applications. Proceedings of the IEEE, Vol. 63, Dec. 1975; pp. 1692-1716

[194] Widrow, B.; McCool, J.M.; Larimoore, M.G.; Johnson, C.R.: Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter. Proceedings of the IEEE, Vol. 64 Aug. 1976, pp. 1151-1162

[195] Wiener, N.: Extrapolation, Interpolation and Smoothing of Stationary Time Series. MIT Press, 1949, pp. 129-139

[196] Wu, Y.S.: Unterdrückung von akustischen Umgebungsgeräuschen mit adaptiven rekursiven Digitalfiltern. Dissertation an der ETH Zürich, 1984

[197] Yang, R.; Haavisto, P.: An improved Noise Compensation Algorithm for Speech recognition in Noise. IEEE Proc. ICASSP ’96, pp. 49-52

[198] Yasmin, A., Fieguth, P., Deng, Li.: Speech Enhancement using Voice Source Models. ICASSP 99, Vol 6. CD-ROM

[199] Zeklinski, R.: Ein Geräuschreduktionssystem mit zweidimensionaler Mikrofongruppe und LMS-gesteuerter adaptiver Nachfilterung. Fortschritte der Akustik, DAGA 91, S. 893-896

[200] Zelinksi, R.: A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms. Proc. ICASSP 88, pp. 2578-2581

[201] Zelinski, R.: Ein Geräuschreduktionssystem mit zweidimensionaler Mikrofongruppe und nachgeschalteter adaptiver Wiener Filterung. Aachener Symposium für Signaltheorie, pp. 372-375, Springer-Verlag, 1987

[202] Zwicker, E.: Psychoacoustics. Springer-Verlag, Berlin-Heidelberg, 1990

[203] Zwicker, E.: Subjektive und objektive Dauer von Schallimpulsen und Schallpausen. Acoustice, Band 22 (1970), S. 214-218

[204] Zwicker, E.; Flottorp, G.; Stevens, S.S: Critical Bandwidth in loudness summation. J. Acoust. Soc. Am., Band 29 (1957), pp. 548-557

Lebenslauf

07.1969	geboren in Teterow, Mecklenburg-Vorpommern
07.1988	Abitur, allgemeine Hochschulreife
09.1988	Wehrdienst
10.1990	Ingenieurhochschule Mittweida: Studium Informationstechnik, Vordiplom
09.1992	Universität Kaiserslautern: Hauptstudium Elektrotechnik (Nachrichtentechnik)
03.1993	Michigan State University (U.S.A.): Studienarbeit Digitale Signalverarbeitung
09.1993	Siemens Automotive L.P, Auburn Hills, MI (U.S.A): Internship
04.1996	BMW AG München, Forschungs- und Ingenieurzentrum: Diplomarbeit
12.1996	Universität Kaiserslautern: Diplom
01.1997	BMW AG München, Forschungs- und Ingenieurzentrum: Doktorarbeit
09.1999	BMW AG München, Entwicklung Elektronik, Strategie und Unternehmens-kommunikation: Projektingenieur
10.1999	BMW AG München, Bereich Kommunikations- und Informationssysteme: Projektleitung Sprachverarbeitung
04.2000	BMW AG München, Forschungs- und Ingenieurzentrum, Abteilung Mensch-Maschine-Schnittstelle: Gruppen- und Entwicklungsleitung Sprachverar-beitung